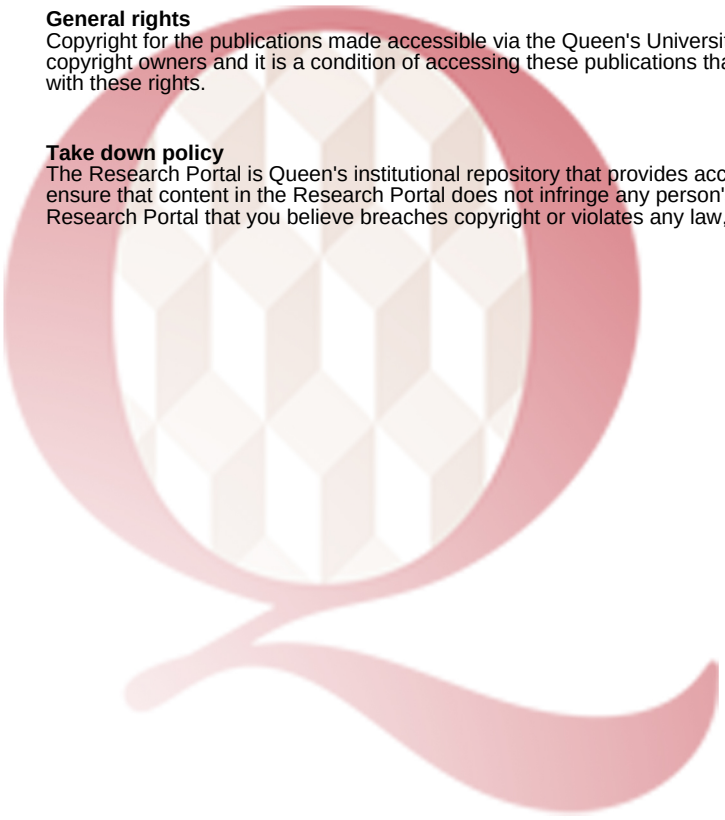
Cross-age tutoring in kindergarten and elementary school settings: A systematic review and meta-analysis

Shenderovich, Y., Thurston, A., & Miller, S. (2016). Cross-age tutoring in kindergarten and elementary school settings: A systematic review and meta-analysis. International Journal of Educational Research, 76, 190-210. DOI: 10.1016/j.ijer.2015.03.007

Published in:
International Journal of Educational Research

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal



A Systematic Review and Meta-Analysis of Cross-Age Tutoring in Kindergarten and Elementary School Settings

Yulia Shenderovich
Institute of Criminology, Sidgwick Avenue, Cambridge, UK

Allen Thurston
Centre for Effective Education, Queen's University Belfast, Northern Ireland, UK

Sarah Miller
Centre for Effective Education, Queen's University Belfast, Northern Ireland, UK

Author note: Correspondence concerning this article should be addressed to Yulia Shenderovich, Institute of Criminology, Sidgwick Avenue, Cambridge, CB3 9DA, UK. E-mail: y.shenderovich@gmail.com. Phone: (+44) 07774955977

Abstract

This systematic review summarizes effects of peer tutoring delivered by non-professional tutors, such as classmates, older children and adult community volunteers, to children between 5 and 11 years old. Inclusion criteria for the review included tutoring studies with a randomized controlled trial design, reliable measures of academic outcomes and a duration of at least 12 weeks.. Searches of electronic databases, previous reviews, and contacts with researchers yielded 11,564 titles. After screening, 15 studies were included in the analysis. Cross-age tutoring showed small significant effects for tutees on the composite measure of reading (g=0.18, 95% CI: 0.08, 0.27, N=8,251), decoding skills (g=0.29, 95% CI: 0.13, 0.44, N=7,081), and reading comprehension (g=0.11, 95% CI: 0.01, 0.21, N=6,945). No significant effects were detected for other reading sub-skills or for mathematics. The benefits to tutees of non-professional cross-age peer tutoring can be given a positive, but weak recommendation. Effect Sizes were modest and in the range -0.02 to 0.29. Questions regarding heterogeneity of effects, study limitations, lack of cost information and the relatively few number of studies that have used a randomized controlled trial design means that the evidence base is not as strong as it could be. Subgroup analyses of included studies indicated that highly-structured reading programs were of more benefit than those that were loosely-structured. Large-scale replication trials using factorial design, process evaluations, reliable outcome measures and logic models are needed to better understand under what conditions, and for whom, cross-age non-professional peer tutoring may be most effective.
Keywords: tutoring, systematic review, literacy, peer learning, volunteer effectiveness, peer tutoring, cooperative learning

Cross-Age Tutoring: A Systematic Review and Meta-Analysis of Cross-Age Tutoring in Kindergarten and Elementary School Settings

Individualized tutoring is considered to be one of the most effective ways to promote improved educational outcomes (Bloom, 1984; Elbaum, 2000; Ireson, 2006). Non-professional peer tutors, such as classmates, community volunteers and university students deliver volunteer peer tutoring programs with reduced costs compared to professional teachers. Most forms of volunteer tutoring are forms of peer tutoring. These tutors are peers by virtue of the fact that they are close in age to the tutees (in the case of school or university student volunteers), or close in terms of background and demography and spatial proximity (in the case of community volunteers who share the living, leisure and local geographical environment with tutees). Therefore, we propose that the term 'peer' tutors should not be restrictive in terms of its description of same age, same school partnerships and we take a wider, inter-generational view of what constitutes a peer within a community. Tutoring by peers/community volunteers has been reported as an effective intervention for improving academic and attitudinal outcomes among school-aged children (Medway, 1995; Higgins, Katsipataki, Kokotsaki,Coleman, Major & Coe, 2013). Conversely, several studies have found null or negative effects for non-professional tutoring on academic results of tutees (Jensen, 1991; McKinney, 1995; Ritter, 2000). Therefore, there is need for a systematic review of such tutoring programmes to assess what high quality, high integrity studies report in terms of the efficacy of peer tutoring.

Theoretical background

There is no single dominant theory of change for peer tutoring. Students are expected to improve academic outcomes through elaborating thoughts in the tutoring process, thus cooperatively constructing knowledge within the so-called *zone of proximal development* (ZPD). The ZPD is loosely defined as the distance between child's independent level of

problem solving and the level of problem solving under the guidance of a more advanced peer or an adult (Vygotsky, 1978; Chi, Bassok, Lewis, Reimann & Glaser, 1989; Webb, 1989). In this manner peer tutoring is often reported as being a form of cooperative learning (Pesci, 2015). Peer tutoring can provide students with timely feedback (Bloom, 1984; Merrill, Reiser, Merrill & Landes 1992), increased time on task (Delquadri, Greenwood, Whorton, Carta & Hall,, 1986) and more appropriate pacing (Shanahan, 1998). Tutoring programs are expected to improve socio-emotional outcomes, such as self-efficacy (Elliott, Arthurs & Williams, 2000), self-confidence (Margolis, 2005), and child's confidence in the academic subject tutored (Koh, Sanders & Meyer, 2012). Peer tutoring could help improve social ties between tutees and tutors (Goodlad & Hirst, 1989), strengthen children's attachment to the school and improve attendance (Pridmore, Stephens & Stephens, 2000). Many authors have also suggested that tutors can serve as role models for the tutees (Allen, 1976; Erickson, 1987; Potter, 1994; Topping & Hill, 1995). In this way, peer tutoring by non-professional educators is expected to be qualitatively different from tutoring delivered by professionals and employed teaching staff.

Ongoing programs

In the USA since the late 1990s America Reads Challenge has mobilized tens of thousands of college students as volunteer reading tutors for children in Kindergarten through Third Grade (Fitzgerald, 2001). In this context, several manualized programs were developed, such as Book Buddies (Meier & Invernizzi, 2001), which involves 45-minute biweekly sessions consisting of rereading a familiar book, word studies, writing, and reading a new book. In India, a programme called India Reads is managed by the largest educational non-governmental organization, Pratham. The programme is reported to enable communities to mobilize and train volunteers to work in schools both during and after school hours. The initiative involves nearly 450,000 community volunteers acting as tutors using techniques

described in programme manuals (Poverty Action Lab, 2009). Other programs have less informal structures for tutoring interactions. The UK literacy charity Beanstalk connected adult community volunteer tutors with 6,400 primary school children in England during the 2011-2012 academic year. It provided community volunteers general guidance, such as "Use open-ended sentences to encourage conversation" and "Be generous with your praise" (Beanstalk, 2013).

Most reports available in English have described tutoring programs in high-income English-speaking countries, such as USA, UK and Australia, but there are also reports of similar projects in other countries, such as China, India, Jamaica, Lithuania, South Africa, Tanzania and Thailand (Goodlad, 1995, 1998). Banerjee and Duflo (2011, pp. 85-86) reported that tutoring programs involving community volunteers are currently being tested in Ghana, with plans for similar programs drafted in Senegal and Mali.

Existing studies and reviews

Following a number of narrative reviews (Rosenshine & Furst, 1969; Devin-Sheehan, Feldman & Allen, 1976), Hartley (1977) carried out the first meta-analysis, identified by this review, of peer tutoring studies in mathematics with child tutors, finding a mean Cohen's *d* of 0.6. The widely cited Cohen, Kulik and Kulik (1982) review examined 65 randomized and matched studies based in elementary and secondary schools with schoolchildren as tutors. It reported significant overall Cohen's *d Effect Sizes* of 0.29 for reading (95% CI 0.17, 0.41) and significant *Effect Sizes* of 0.6 (95% CI 0.29, 0.91) for mathematics. However, Rohrbeck, Ginsburg-Block, Fantuzzo and Miller (2003) reported that older meta-analyses may have serious methodological limitations, such as 'lax' and 'non-transparent' study inclusion criteria. More recent reviews (Wasik & Slavin, 1993; Shanahan, 1998; Wasik, 1998). Elbaum, Vaughn, Hughes & Moody, 2000) looked at one-to-one tutoring undertaken by adults, including professional tutors. It was reported that, *"college students and trained,*

reliable adult community volunteers were able to provide significant help to struggling readers" (Elbaum et al., 2000, p. 616).

More recently, Slavin and colleagues (Slavin & Lake 2008a; Slavin, Lake, Chambers, Cheung & Davis, 2009a; Slavin, Lake, Cheung & Davis 2009b; Slavin, Lake, Chambers, Cheung & Davis,2009c; Slavin, Lake, Davis & Madden, 2010; Slavin & Madden, 2011) carried out large Best Evidence Encyclopedia syntheses of various reading programs in Kindergarten to Grade 5. The reviews reported significant standardized mean difference *Effect Sizes* of 0.26 for cross-age tutoring. Leung, Marsh and Craven (2005) conducted a meta-analysis of 68 published studies, in which children and university students were tutors. It was reported that there were significant *Effect Sizes* of 0.65 for overall academic achievement (95% CI: 0.59, 0.71) and 0.88 for self-concept (95% CI: 0.69, 1.07). In contrast Torgerson and King (2002) and Ritter, Denny, Albin, Barnett and Blankenship (2006) focused on randomized controlled trials (RCTs) with adult non-professional tutors. Torgerson and King (2002) summarized four trials, finding a mean *Effect Size* of 0.19 that was not statistically significant (95% CI: -0.31, 0.68). Ritter et al. included 21 USA based studies, finding a significant mean *Effect Size* of 0.3 (95% CI: 0.18, 0.42) for the composite measure of reading and a non-significant mean *Effect Size* of 0.27 (95% CI: -0.18, 0.72) for mathematics. A recent review of 76 randomized experiments in education conducted in low and middle income countries found an average effect of 0.10 for community volunteer teaching (McEwan, 2013). These *Effect Size* estimates are lower than those reported by Leung et al. (2005). Thus, results of previous meta-analyses ranged from null to small/medium positive significant effects.

Given the wide diversity of effects identified in previous research, the current review was deemed necessary to systematically identify randomized studies in this area, critically appraise available evidence and provide a more precise estimate of the effect of tutoring on

academic outcomes, including the most recent research evidence. Given the wide use of tutoring programs, this review is needed to make suggestions for teaching as well as inform possible directions for future research.

Method

Inclusion criteria

To develop inclusion criteria for the review and ensure that only studies with high methodological rigour were included, current criteria published by What Works Clearinghouse (2010), Cochrane Collaboration (Higgins & Green, 2011) and Best Evidence Encyclopedia (2013) were examined. After close examination and discussion within the review team, a full list of inclusion criteria for this review was developed as follows.

Sample size included at least two classrooms per treatment group. Contextual factors in education research are important (McCartney & Ellis, 2008). In small-scale studies, intervention effects are likely to have confounds with particular schools, classes or teachers, dramatically limiting generalizability of the results. There will be some common attributes of the 'cluster' and there is a danger in single classroom/context studies that that the strength of common or 'clustered' attribute may be more powerful than the effects of the intervention e.g. teacher quality, school quality or socio-economic status of participants (Slavin & Smith, 2009). Therefore, in agreement with What Works Clearinghouse guidelines (What Works Clearinghouse, 2008), studies with only one classroom per treatment were not included due to the risk of single context clustering effects biasing reported outcomes.

Randomization was used to assign to treatment or control condition. Randomized controlled trials (RCTs) are studies, in which participants, or groups of participants, are randomly assigned to experimental and control groups. The experimental participants receive treatment, while control participants receive treatment as usual, an alternative treatment or no treatment at all (Bowling, 2009). Randomized controlled trials are widely

recognized as the most reliable research design to assess the effectiveness of an intervention as they create two equivalent groups and identifying intervention effects (Guyatt, Oxman, Kunz, Falck-Ytter, Vist, Liberati & Schünemann, 2000; Glazerman, Levy & Myers, 2003; Petticrew & Roberts, 2003; Agodini & Dynarski, 2004; Wilde & Hollister, 2007). Although randomized controlled trials and high-quality matched studies may identify similar *Effect Sizes* (Torgerson, 2007), randomized controlled trials and matched studies do not always lead to same conclusions (Heinsman & Shadish, 1996; Glazerman et al., 2002). RCTs tend to report lower Effect Sizes than matched design experiments. This review relies exclusively on studies with an RCT research design so that outcomes were not unduly affected by research design as opposed to quality of tutoring.

Outcome measures did not bias treatment over control condition. The review included studies with measures that were reliable and valid. A measure is inherent to the experimental treatment if it assesses particular skills or concepts that have been taught only to the experimental group. Miller, Maguire and Macdonald (2012) reported that measures described as directly related to the program's goals may be inherent to the treatment and thus bias any comparison in favour of the intervention group. It follows that findings of a study are determined not only by the intervention investigated and the nature of the comparison group, but also by the quality and independence of measures used. Gersten Baker and Lloyd (2000) highlighted that when experimental designing was undertaken in special education, it was important to distinguish experimenter-developed and external measures. This review included studies that used attainment scales of any academic ability in which the reliability and validity of measures could be ascertained, e.g. by issue of a standardized instrument or at least a full description of the psychometric properties of the scale and its scoring being available. Pre-test differences between control and treatments groups had to be reported as non-significant, or with pre-test differences controlled for during analysis.

Outcome measures of academic or socio-emotional ability. Secondary outcomes are outcomes that are not priority of the review, but are important for explaining effects (O'Connor, Green & Higgins, 2008). Tutoring is theorized to rely not only on cognitive, but also socio-emotional outcomes (Robinson, Schofield & Steers-Wentzell, 2005), such as confidence (Koh et al., 2012), self-efficacy (Elliott et al., 2000) and self-confidence (Margolis, 2005). Therefore, although academic outcomes were the primary aim of the review, socio-emotional results, if available, were included as secondary outcomes.

Intervention length was 12 weeks or longer. The review focused on *"practical programs that can be used over extended time periods, not theoretically interesting but impractical procedures that could never be replicated for extended periods"* (Slavin, 2008, p. 11). Consequently, to achieve higher external validity and relevance to school practice, the minimum length for a study to be included in this review was 12 weeks between pre-test and post-test, following Best Evidence Encyclopedia standards (Center for Data-Driven Reform in Education, 2013) on this issue. In contrast very short programs may not lead to forming sustainable habits (Lally, Jaarsveld, Potts & Wardle, 2010).

Nature of tutoring.

1) School-based programs using individualized instruction in dyads or small groups, involving a more academically advanced tutor and one or more less advanced tutees (Medway, 1995; Topping, 1998).

2) Tutor and tutee had fixed roles, i.e. tutoring was non-reciprocal and tutors/tutees remained in those roles for the duration of the programme.

3) Tutoring was delivered by classmates/older students, parents, university students or other adults (for example community volunteers) acting in a non-professional tutoring role. Paraprofessional and professional teachers, and professional tutors were excluded.[1]

4) Tutoring took place in a face-to-face setting (this was used as a inclusion criteria as the differences between face-to-face and on-line tutoring not yet fully explored in the research literature).

5) Tutoring was carried out within the school context of the tutee.

6) The recipients of the tutoring were tutees in a kindergarten/primary/elementary school setting, which corresponds to the age bracket of five to eleven years old.

7) Tutoring had an academic focus in any subject area.

8) Outcome measures included attainment tests and information was provided that allowed *Effect Sizes* to be calculated from the reported data.

9) Intervention tested tutoring on its own without significant additional components, such as scholarships.

10) The duration of the tutoring intervention was not less than 12 weeks long.

Search strategy for identification of studies

Given the spread of published educational intervention research over many resources (Newman, 2003; Glanville & Paisley, 2010), a wide range of databases were identified to reduce the possibility of missing studies. In addition to databases, organization's websites, bibliographies of key studies, literature reviews and meta-analyses were analysed for review titles. Modifications of the search string *tutor* AND (peer* OR cross-age OR volunteer*) AND (evaluation* OR program* OR experiment* OR random*) NOT technolog** was used on ASSIA, Australian Education Index, British Education Index, ERIC, International Bibliography of Social Sciences, JSTOR, PsycINFO, PRISMA, ProQuest Dissertations &

[1] To distinguish volunteers and paraprofessionals, this review considered tutors to be volunteers if they received no payment at all or if they were only reimbursed for travel to the school (Lee et al., 2010) and other participation costs incurred (Cabezas et al., 2011).

Theses, Web of Knowledge, Social Services Abstracts, and Sociological Abstracts. In addition, 104 researchers were contacted by email to identify unpublished studies. It should be noted that use of * at the end or in the middle of a word will return searches of all letter strings/spellings that are contained in the string. For example randomi*ed would return all search items with spelling of both randomized (USA spelling) and randomised (UK spelling).

Insert Figure 1 about here

Data presented in Figure 1 shows the flow diagram of identification and screening of studies. A total of 11564 titles were retrieved through the review searches. Citations were imported into Microsoft Excel, which was used to remove duplicated records, leaving 10,910 unique titles and abstracts. Initial screening titles and abstracts left 183 studies for further review. Full texts of the 183 studies were obtained and assessed for eligibility by first author (Shenderovich). A randomly selected 20% of studies were further screened by both first and second authors (Thurston) with no disagreements. Shenderovich and Thurston also examined the full list of titles to discuss any studies that caused dubiety as to whether further screening would be required and made decisions in each case. Fifteen studies (reporting data from 16 cohorts of participants) fulfilled all inclusion criteria as determined by two authors. All studies were fully coded by the first author, and half were blind double-coded by both reviewers. The other half of included studies was checked by the second reviewer for coding accuracy and to ensure inclusion criteria were met.

Effect Size calculations

To determine if tutoring had greater effect in any area of reading sub-skills, reading outcomes were categorized under the following categories for separate meta-analyses: comprehension, fluency, decoding, writing and overall reading ability, using the approach adopted by Ritter et al. (2006) in their review. As mathematics outcomes are categorically different from reading outcomes, reading and mathematics outcomes were maintained as separate variables. In cases where a more than one measure within a study assessed the same construct, to make sure that no study was unduly weighted in the analysis of *Effect Sizes* and their confidence intervals were averaged (Becker, Hedges & Pigott., 2004), assuming a correlation of 0.5 between related scores (Borenstein, Hedges, Higgins & Rothstein, 2009).

Analyses were carried out using Comprehensive Meta-Analysis software, version 2 (Biostat Englewood, NJ). Standardized mean difference (Cohen's *d*) is the appropriate *Effect Size* metric to contrast two groups on continuous variables, such as test performance (Lipsey & Wilson, 2001). Standardized mean difference is calculated as difference in mean outcomes between groups divided by pooled standard deviation of outcome among participants. *Effect Sizes* and confidence intervals were divided by Hedges's approximation (Hedges & Olkin, 1985; Lipsey & Wilson, 2001). Given the diversity of tutoring programs, random-effects model was pre-selected in the review protocol to make studies more equally weighted (Sterne, Egger & Smith, 2008) and results more generalizable (Field, 2001). Manuscript authors were contacted directly if any missing information was needed to calculate *Effect Sizes*.

In educational research it is common to assign groups of children, such as classes or schools, to treatment and control groups (Boruch, May, Turner, Lavenberg, Petrosino, De Moya & Foley, 2004; Campbell, Elbourne & Altman, 2004). The effective sample size in a cluster-randomized trial is the original sample size divided by the "design effect", which equals 1+(M−1)*ICC, where M is the average cluster size and ICC is the intra-cluster correlation coefficient (Higgins et al., 2008). ICC adjustment was applied for the Elliott et al. (2000) study, the only included cluster-randomized trial, using ICC of 0.15, the value suggested by a recent compilation of research on intraclass correlation values of academic achievement in the USA (Hedges & Hedberg, 2007).

Results

Description of the included studies

As described in Table 1, eleven of the investigations were carried out in USA, four in the UK and one in Chile. The majority of tutoring programs focused on low-achieving children, indicated either by their classroom teacher or test assessment. In respect to external validity,

it is important to point out that the majority of studies recruited what appeared to be a convenience sampling of classrooms and schools, and are therefore not necessarily generalizable to other settings. However, some studies used representative samples, either of local schools (Miller & Connolly, 2012) or of the tutoring programme's participants (Lee, Morrow-Howell, Jonson-Reid & McCrary, 2010). All studies focused on schools with disadvantaged socio-economic profiles. Several programs targeted one age group (Pullen, Lan & Monaghan, 2004; Allor & McCathren, 2004 – Group (Gr) 1, Cabezas, Cuesta & Gallego, 2011 – Gr 4), while others included a variety of primary school grades (Ritter, 2000 – Gr 2-5, Lee et al., 2010 – Gr 2-3).

Study sizes ranged from small-scale trials with 42 (Rimm-Kaufmann, Kagan & Byers, 1998) and 47 children (Pullen et al., 2004), to large-scale studies with 734 (Miller & Connolly, 2012), 883 (Lee et al., 2010) and 6136 children (Cabezas et al., 2011) enrolled respectively. In total studies involved 9484 participants. Following the approach of Best Evidence Encyclopaedia, this review defines large studies as those with greater than 250 participants (Slavin, 2009). Five included studies with samples over 250 looked at on-going programs (Experience Corps, West Philadelphia Tutoring Project, Time to Read, Servicio País en Educación) in multiple locations and, thus, were effectiveness—as opposed to efficacy—studies (Haynes, 1999; Flay, 1986, 2005).

Most included studies focused on reading, and two studies involved tutoring in mathematics. Ham (1977) assessed the "halo effect" of tutoring in reading on achievement in mathematics. The observed emphasis on reading focused studies could be reflective of the importance of reading in primary school, as well as of the more complex nature of designing tutoring procedures in mathematics (Topping, 2004). Studies identified by this review did not target any other academic subjects.

Two cohorts included in the review utilized older schoolchildren as tutors (Jensen, 1991; Policy Studies Associates, 2007), and fourteen investigated tutoring by adults (eight of them involved adult community volunteers, and six with university student volunteers). All studies except one involved English-language instruction (Cabezas et al., 2011 studied reading in Spanish language in Chile). In addition to tackling outcomes of primary school tutees, some of programs aimed to improve achievement of tutees who were school or university students (Policy Studies Associates, 2007) or to contribute to social wellbeing of older tutors (Lee et al., 2010).

Seven studies examined programs that prescribed specific tutoring lessons and materials or specified time allocated for various activities. This review characterizes such programs as "highly structured" – incorporating standardization by precise activities or by functions and processes (Baumann, 1991; Backer, 2001). More structured programs also had more extensive tutor training. For instance, Pullen et al. (2004) provided university student volunteers with step-by-step lesson guides, and the tutoring sessions were observed by supervisors. On the other hand, nine studies provided general advice to tutors and are therefore classified as "loosely structured". For example, in Northern Ireland the Time to Read program, evaluated by Miller, Connolly, and Maguire (2012), adult community volunteers did not receive a pre-set tutoring session structure. In Baker et al. (2000), adult community volunteers were *"provided with a broad framework to use during sessions, rather than specific techniques"* (p. 497). Similarly, in the Ritter (2000) evaluation of West Philadelphia Tutoring Project, tutors (University of Pennsylvania volunteer students) had only general guidance on working with their tutees and curriculum guides were only provided in some of the participating schools. There was no structured process evaluation, but anecdotal reports suggested that during sessions tutors helped pupils with homework tasks or made up their own exercises in reading and mathematics.

Insert Table 1 about here

Description of excluded studies

Most studies were excluded due to lack of randomization. In addition, to examine sustainability, a minimum of 12 weeks length was set for inclusion, as discussed above, which left out several otherwise eligible studies. For instance Spörer, Brunstein and Kieschke (2009), randomized 210 elementary school children from 4 classes in a medium-sized German town to four groups: instructor-guided small groups; direct instruction followed by reciprocal tutoring; a mix of direct instruction and reciprocal tutoring; and a no-intervention control group. However, the study only lasted seven weeks. In addition, several studies were excluded because of a lack of eligible comparison groups.

In another excluded paper, an unpublished study based in migrant schools in Beijing, China (Li et al., 2010), all study groups were paid for grades, and, in addition, a third of the 850 students received tutoring from classmates and a third tutoring from classmates, plus a parental communication intervention. Thus, there was not a tutoring only group where no payment was made available. It was reported that tutoring and pay showed an *Effect Size* of 0.14 on reading and the group with tutoring and pay plus parental communication had an *Effect Size* of 0.2. Another study (Banerjee, Banerji, Duflo, Glennerster & Khemani, 2010) describes a set of interventions evaluated in 65 randomly assigned villages in India in 2005. Similarly, none of the interventions tested tutoring on its own, so the study was not included. All three interventions involved sharing information on educational resources with communities through small-group discussions. A second intervention also included offering communities testing tools to assess children's reading and mathematics results, and the third facilitated community volunteer tutors providing afterschool reading.

Overall effects

The review suggested small (as defined in Cohen, 1988) statistically significant positive effects, with high heterogeneity, of cross-age tutoring programs on reading overall, as well on decoding and comprehension skills, while outcomes on other reading measures and mathematics were non-significant. The high heterogeneity of findings for many of the non-reciprocal tutoring outcomes indicates that the studies, populations and interventions included are diverse.

Outcome measures were grouped into seven categories, following the example of the Ritter et al. (2006) systematic review:

- Composite measure of reading: measure combining all reading measures available in each study (see Forest plot in Figure 2)
- Overall reading: overall batteries on reading achievement tests
- Decoding: in this category, the review included subtests on decoding of words and knowledge of words, consonant sounds, short vowels, digraphs and combinations, sight words, and non-word decoding
- Comprehension: in this category, the review included reading comprehension subtests
- Fluency: in this category, the review included fluency subtests
- Writing: in this category, the review included achievement tests on assessment of student writing
- Mathematics: in this category, the review included measures assessing mathematics outcomes

These seven categories covered the reported attainment measures of all included studies and therefore forms an all-inclusive set of outcome descriptors. Figure 2 shows the composite measure of reading, with upper and lower *Effect Sizes* for the battery of tests reported by each manuscript.

Insert Figure 2 about here

Homogeneity analysis

Table 2 lists several measures of homogeneity. *Q* represents a standardized measure of total variation, and *df*, the expected variation. Thus *Q* minus *df* is the excess variation. The *Q* statistic and its *p*-value are a test of significance of the viability of the null hypothesis of zero true dispersion. I^2 is the percentage of the dispersion that is real and not due to sampling error. Higgins, Thompson, Deeks & Altman, (2003) tentatively suggest that I^2 values of 25%, 50%, and 75% are respectively low, moderate, and high, with about a quarter of meta-analyses having I^2 over 50%. Finally, T^2 is the variance and T , the standard deviation of true effects, measured on the same scale as effects. The level of heterogeneity for decoding, fluency and composite measure of reading was high. Nevertheless, Ioannidis, Patsopoulos and Rothstein (2008) suggest that overall meta-analysis is usually desirable, even with high statistical heterogeneity. Although statistical homogeneity tests are weak and not very precise (Ioannidis et al., 2007; Thorlund Imberger, Johnston, Walsh, Awad, Thabane, Gluud, Devereaux & Wetterslev, 2012), statistical heterogeneity can be a useful tool (Berlin, 1995) as it points to the presence of clinical or methodological diversity, or both (Deeks, Higgins & Altman, 2011).

Insert Table 2 about here

Sensitivity analysis

Sensitivity analysis is necessary to assess potential bias that may be associated with individual *Effect Sizes* and distort the aggregated effects (Hedges & Olkin, 1985). "One Study Removed" analysis allows to assess if any single study has disproportionate influence. In this set of studies, several very large samples are present. In particular the large sample (N=4,903) in Cabezas et al. (2011) made up 59% of all reading studies' participants. Using a random effects model, all estimates with one study removed fell inside the 95% confidence

interval of the overall estimate with all available studies. Therefore no study was found to have an excessive influence on results.

Publication bias

Five of the included studies have not been published in academic journals. Three were dissertations and two were reports. Non-significant or negative results, especially in small-sample studies, are often not submitted or not accepted for publication, although they may be of equal quality as published work (Iyenger & Greenhouse, 1988; Hopewell, Loudon, Clarke, Oxman & Dickersin, 2009). To assess the possibility of publication bias, the "trim and fill" procedure (Duval & Tweedie, 2000) was conducted for each outcome to identify and correct funnel plot asymmetry (see Figure 3 for composite measure of reading funnel plot). The "trim and fill" procedure for the composite measure of reading did not indicate any missing studies. However, there was an indication of studies missing to the left of mean effect sizes for the overall reading ability, comprehension, decoding, and mathematics measures, suggesting possible publication bias. The impact of publication bias still may be trivial as at least 8-10 studies are required for trim-and-fill test to have sufficient power (Sutton, Duval, Tweedie, Abrams & Jones, 2000a, 2000b). In addition, Egger's regression testing asymmetry of the funnel plot was not significant ($p>0.05$) for any measure, indicating low risk of publication bias, although the small number of studies does not allow for definitive conclusions.

Insert Figure 3 about here

Moderator analyses and meta-regressions

Several program features were examined through subgroup analyses and meta-regressions. Grouping of studies was used to assess the possibility of varying reading outcomes of different types of programs to analyse possible sources of heterogeneity (see

Table 3 for a summary). Mixed effects analysis was used, meaning that random-effects model is used within groups and fixed effects across subgroups with pooled estimates of T^2. Studies were grouped by the variable of interest, and subgroup effects were compared using significance of *Q* to see if *Effect Sizes* between groups were statistically different.

Study size. Eleven studies had samples of 30 to 157 children, and were coded as "small", while five studies with samples of 328 to 4903 were coded as "large". Difference between two groups was statistically significant for Composite measure of reading (p=0.008) and Decoding (p<0.001), with larger studies showing significantly smaller effects than smaller studies. This is a common feature when reporting data in systematic review and comparing studies. Similarly to previous studies, there were much higher levels of heterogeneity among smaller studies (Q=41.176, df=10, p=0.000, I^2=75.714) than among larger studies (Q=3.714, df=4, p=0.446, I^2=0.000). Smaller studies are subject to higher sampling variation (Higgins & Altman, 2008) and have low statistical power, increasing likelihood of a false positive result (Christley, 2010). Larger studies produce more precise estimates as they are generally correctly powered to detect effects (Ginsburg-Block, Rohrbeck & Fantuzzo, 2006). Method of moments meta-regression suggests no significant correlation between study size and composite measure of reading (p_{slope}=0.315).

Tutoring structure. Highly structured programs (9 studies, g=0.33, 95% CI: 0.14, 0.52, N=1,388) had a significant advantage over programs with low structures in place (7 studies, g=0.08, 95% CI: -0.01, 0.16, N=6,863) on the Composite measure of reading outcome. Comparing groups with the *Q*-test (Borenstein, Hedges, Higgins & Rothstein, 2009, p. 178), Q=5.903, p=0.02, thus *Q* is statistically significant, and *Effect Size* is related to the level of structure.

Type of tutor. Subgroup differences by type of tutor comparing tutors who were university students, adult community volunteers or peer tutors did not indicate significant differences in random effect analysis.

Publication status. Subgroup differences depending on publication status being published or unpublished report or thesis did not indicate significant differences in random effect analysis.

Insert Table 3 about here

Amount of tutoring. Method of moments meta-regression examines differences in the effect of tutoring on composite measures of reading, depending on 'dose' of tutoring, as measured by the number of tutoring hours. Amount of tutoring did not give a good explanation of effectiveness of tutoring in included studies (p_{slope}=0.584).

Social, self-concept and behavioural outcomes

Few studies included in this review tested non-academic outcomes alongside academic skills. Due to their diversity and small number, non-academic results were not meta-analysed but are summarized in Table 4, and all were non-significant except one.

Insert Table 4 about here

The quality of evidence

Littell, Corcoran and Pillai (2008, p. 72) propose that *"Even when a review is limited to randomized controlled trials, a deeper assessment is needed to judge variations in quality of those studies that may be associated with bias."* This is particularly important because randomized controlled trials in school and educational settings are reported to have lower quality than in healthcare (Torgerson, Torgerson, Birks & Porthouse, 2005). Assessments of

domains of bias specified in Cochrane Collaboration Risk of Bias Tool (Higgins & Altman, 2008) are outlined below. As reported in Table 5, the included studies did not display many areas of potential bias.

Insert Table 5 about here.

Selection bias. Only four studies specified their approach to generation of randomization sequence, and all four used computer-generated sequences. Two studies, Loenen (1989) and Ritter (2000) discussed practical challenges surrounding gaining cooperation from schools for randomization. Therefore, it is not possible to rule our selection bias as a contributing factor to effects in some studies.

Performance and detection bias. Although blinding of study participants and intervention personnel (such as class teachers and tutors) is not possible in a tutoring intervention, it may be possible to blind the assessors. Six if the studies did this. Rimm-Kaufmann, Kagan and Byers, (1998) reported that classroom teachers were blinded to which children were assigned to the control group.

Attrition bias. The studies described a wide range of attrition levels, some as high as 35%. There was no standard approach to intention to treat analysis and so it was not possible to assess attrition risk in a quantifiable manner.

Reporting bias. The presence of differences between reported and unreported findings could not be assessed due to lack of study protocols

Other biases. 1) There were significant pre-treatment (baseline) differences between treatment and control groups (either due to chance or problems with randomization) in two studies (Jensen, 1991; Pullen, Lane & Monaghan, 2004), but it was reported that differences were accounted for in ANCOVA analyses.

2) There was a lack of long-term follow up measurements in the included studies. A possible explanation for this may be sue to ethical and practical difficulties of having a no-intervention control group in schools. Only Policy Studies Associates (2007) and Elliott et al. (2000) studies had follow-up assessments. Thus the review is primarily based on post-test (tests at the end of interventions) rather than on follow-up measures. Longevity of change was therefore difficult to assess.

3) Five large studies used multilevel modelling to account for classroom and school effects. However, smaller studies did not adjust for clustering effects within classrooms and schools, and as Miller & Connolly (2012, p. 12) note, "clustered nature of data" is present when children come from the same classrooms and schools, violating statistical assumptions of independence.

Discussion

Whilst publication bias was not apparent, evidence presented by the review must be viewed with caution due to high heterogeneity, quality limitations and small number of included studies. The review suggested that tutoring programs had small positive effects on combined measures of reading as well as specifically on decoding and comprehension. However, Chall's synthesis of theories of reading concludes that both decoding and fluency skills are necessary for comprehension skills to develop (Chall, 1989). One explanation is that decoding and comprehension measures had more eligible larger and correctly powered studies included in the synthesis, and thus the meta-analyses for these measures had more power to detect effects (Borenstein, Hedges, Higgins, and Rothstein, 2009).

In-line with previous reviews on tutoring (Fitz-Gibbon, 1977; Palincsar & Brown, 1989; Wasik & Slavin 1993; Ginsburg-Block, 2006; Ritter, 2009; Ewan 2013), studies with a pre-set structure of tutoring report greater *Effect Sizes*. This could support the idea that *"open-ended discussions and explanations are problematic, confusing and ineffective"* (Fuchs et al.

2001, p. 16). Non-trained tutor behaviours have been reported to use 'knowledge-telling' rather than 'knowledge-building' explanations (Roscoe & Chi, 2007). However, findings of subgroup analyses are observational and should be treated with caution as we cannot account for potential confounders. For example, it is also possible that more structured programs were better organized in other respects, such as better tutor training. Moderator analyses suggested that using different types of reading tutors, depending on who is available in the given community, could produce similar results, if a structured tutoring program was established. However, the number of studies is small, and only two eligible studies with child tutors were identified.

Based on meta-regression results, there was no difference in reading outcomes by dose of tutoring, as measured by number of hours. It should be noted that meta-regressions have very weak statistical power a low number of studies. Regarding this apparent lack of dose-response relationship in tutoring, the findings of this review are in line with results of recent large-scale randomized trial of peer tutoring study in Scotland, The Fife Peer Learning Trial (Tymms, Merrell, Thurston, Andor, Topping & Miller, 2011). A no-intervention control group was absent, and the different groups served as controls to each other (e.g. reading tutoring children served as controls for mathematics and vice-versa), so the study was not included in this review. The study was a large-scale district-wide effectiveness trial involving two-15 week tutoring periods spread out over two years (129 elementary schools, nearly 9,000 pupils). The factorial design examined effects of intensity (once per week against three times per week), cross-age (10 year olds tutoring 8-year olds) against same-age tutoring (8-year olds) and tutoring in maths only, reading only and both reading and maths. HLM analysis indicated that intensity did not have a significant effect on outcomes in Performance Indicators in Primary Schools standardized tests, but that *Effect Sizes* for cross-age tutoring were significantly greater than for same-age tutoring.(0.25 as compared to 0.02).

On the other hand, Vadasy, Jenkins, Antil, Phillips and Pool (1997) compared a group of paraprofessional tutors who came to each session and tutored the full amount of time to a group who did not follow time commitments as closely. The study found much higher *Effect Sizes* for tutees whose tutors attended regularly, suggesting that quantity of tutoring may have an impact on student outcomes. However, it should be noted that the study had a small sample of 20 students. Similarly, in Lee et al. (2010) reported gains were slightly stronger (*Effect Size* 0.01-0.04) on three out of four decoding measures for students who received at least 35 tutoring sessions. However, it is possible that the Fife Peer Learning Project gives better comparability as students received fewer sessions by design and findings were unlikely to be biased by clustering effects of the quality of implementation.

There was not a significant correlation between study size and *Effect Size*, but the five large tutoring studies had significantly lower effects than the smaller studies. Thus, the large studies seemed to disagree with the smaller ones. Four out of five of the largest cross-age tutoring studies also had low-structure sessions, so differences could have been an artefact of low structure of sessions in the large studies. Still, this difference could point to super-realization bias as smaller studies are often very closely overseen by researchers (Cronbach, 1980). LeLorier, Gregoire, Benhaddad, Lapierre and Derderian(1997) reviewed clinical medical interventions and reported that outcomes larger studies (12 of 1000 patients or more) were not predicted accurately 35% of the time by earlier meta-analyses on the same topics. Based on included studies, it appears likely that *"the larger studies tend to be those conducted with more methodological rigour, or conducted in circumstances more typical of the use of the intervention in practice"* (Sterne et al., 2008, p. 321), so evidence from large trials needs to be given priority when using systematic reviews to report results that may be genarlizable.

<u>Implications for research</u>

Rigorous study design and methods of reporting needs to be examined closely by the educational academic community. One of the important observations from this review is the need for standardized publication of research protocols. Ideally this should take place prior to research being conducted. It should make particular note of protocols for randomization including any corrections to block or minimise the control and intervention samples. In addition it is vital that data is given on demographics of research participants. Some of the key demographic information about participating children, such as their gender and socioeconomic background, was not reported in detail in the majority of studies. Participant demographic information allows for moderator analyses (Gardner 2006, 2010; Drugli, 2010) to help better understand what works for whom and under which conditions (Hargreaves, 1996). For instance, Cabezas, Cuesta and Gallego (2011), reported that overall program effects were not significant, but subgroup analyses indicated a significant positive impact on reading in low socio-economic status public schools in Bio Bio Region. In addition, the ultimate purpose of interventions are "important gains […] generalized and maintained over time" (Mullen, 2006, p.85). Studies with long-term follow-up are needed (Flay et al., 2005), particularly in mathematics as only two mathematics tutoring programs were identified by the review.

Only Lee (1980) and Ritter (2000) studies discussed matching tutors and tutees, although matching has been described as an important program element by many authors (Wood & Bruner 1976; Reisner et al. 1989; Topping & Whiteley, 1993; University of Barcelona, 2007; Naidoo, 2009).

The implementing organizations also merit more description in future research, given recent evidence suggesting that it can also be very important to student outcomes in educational programs (Bold, Kimenyi, Mwabu, Ng'aga'a, Sandefur, 2013). For example Bold et al. (2013) found that short-term teacher contracts increased student achievement in

Kenya when implemented by non-governmental organization World Vision Kenya, but showed no effects in provinces randomly allocated to government implementation. The researchers explained their findings were potentially due to differences in fidelity of implementation, although fidelity was not formally assessed. The research team concluded that the influence of implementing organization is so significant that even findings from effectiveness studies may not be directly relevant to program implementation in real-world settings if the implementation agent is different from the one researched. Organizations undertaking RCTs might be have *"stronger drive for performance or generally stronger capability"* (Pritchett & Sandefur, p.31).

Emphasis on theory of change. Previous reviews discussed that tutoring programs need stronger theoretical grounding (Devin-Sheehan et al., 1976; Rohrbeck et al., 2003). As Miller, Connolly and Maguire (2012, p. 140) point out, *"it remains relatively unknown how or why volunteer mentoring programmes are effective".* Every intervention is based on theories (Weiss, 1997; Bickman, 2000). To be tested, theories can be expressed, for instance, in a logic model (Zief, Lawyer & Maynard, 2006; Cooksy, Gill & Kelly, 2001) or Causal Chain Analysis (Loyalka, Liu, Song, Yi, Huang, Wei, Zhang, Shi, Chu & Rozelle, 2013). In particular, tutoring is theorized to also rely on socio-emotional processes, but *"tutoring programs have placed greatest emphasis on cognitive processing"* (Shanahan, 1998, p. 231). Similarly to previous reviews (e.g., Cohen et al., 1982, Ritter et al., 2009), this systematic review identified few studies measuring socio-emotional outcomes. Developing and testing logic models for peer tutoring programs could also help to distinguish between elements that are essential and variable in the intervention (Craig, Dieppe, Macintyre, Michie, Nazareth & Petticrew, 2008).

Perhaps the best way to compare components of an intervention is within a randomized controlled factorial trial (Deeks et al., 2011). If sufficient sample sizes are recruited, it will be

particularly beneficial to compare several types of tutoring and different types of tutors. Particularly few studies have investigated non-reciprocal tutoring by children. Otherwise, there is danger of being unable to detect how variables such as tutor competence, training, time etc may predict outcomes.

Process evaluation. Even potentially effective programs may fail to improve outcomes due to how treatment was delivered (Dobson 1980; Hawe, Sheil & Riley, 2004; Mihalic, 2004). Process evaluations add crucial insights to study results (Linnan & Steckler, 2002; Lewin, 2009). For instance, the Loenen (1989, p. 310) study involved observations of 30 tutoring sessions and characterized them as *"different from VRH [Volunteer Reading Help charity, currently Beanstalk] presented in the initial VRH training course"*. Topping, Miller, Murray and Conlin (2011) undertook process observations in the Fife Trial and data suggested that "tutoring technique was only partly implemented". Lack of fidelity assessment may produce descriptive ambiguity (Rychetnik, Frommer, Hawe & Shiell, 2002), and result in researchers *"evaluating a program that has not been adequately implemented"* (Basch, Sliepcevich, Gold, Duncan & Kolbe, 1985, p. 316). Process observations can further illuminate the theory of change through testing correlation between implementation variables and attainment (Topping, Thurston, McGavock & Conlin, 2012).

As part of the process evaluation, intervention cost should be recorded and reported as it informs subsequent recommendations about using an intervention, along with the quality of evidence (Guyatt et al., 2008a; Krishnaratne, White & Carpenter, 2013). Resource scarcity is a notorious issue in education, and it is important to record all resources, including personnel and materials required (McEwan, 2012). Although many programs mention that they are less costly then employing professional tutors, only Ham (1977) has given the actual program costs, and Cabezas et al 2011 provided a cost-benefit analysis.

Implications for implementation of tutoring programs

Based on the limited sample of included studies, it appears that using highly structured interactions between tutor and tutee is important. In the West Philadelphia Tutoring Program, Ritter and Maynard (2008) highlighted the lack of tutor training and tutoring session structure to explain the absence of positive effects. Ritter and Maynard also concluded that highly structured tutoring programs are more likely to lead to improved reading. Similar phenomenon was observed in the Fife Peer Learning study, which reported *Effect Sizes* of 0.2-0.25 for highly structured peer tutoring in mathematics (Tymms et al, 2011). The impact of structure shows the important role that an educator has in designing tutoring programs to ensure that interactions maximize the behaviours seen as providing effective learning.

This review included only 16 study cohorts, so any findings must be treated with some degree of caution. Nevertheless, as the lack of statistically significant student improvements on some measures indicated, cross-age tutoring may not always increase academic outcomes as intended. While this review focused on benefits to tutees, some evidence suggested that children benefit was greater when acting in the role of peer tutor (Robinson, 2005). Therefore, this review does not assess the overall benefit of tutoring programs. This is one of the limitations of the review. Although a transparent and rigorous search strategy was employed, study selection and quality appraisal was intentionally set to a level whereby findings may have been generalised to different educational contexts. However, the small number yet wide diversity of eligible studies limits the strengths of conclusions. The authors are currently undertaking a large-scale (128 class) cluster randomized trial of cross-age peer tutoring where the differential benefits to tutors and tutees of tutoring programs will be assessed.

In conclusion there are lessons and messages for both for practitioners and researchers from the review. Practitioners need to be aware that studies are not consistent in the

definitions of “tutoring”, “mentoring” and “volunteering”, so it is important to obtain the specific program descriptions so they are clear about the structure and form/function of interactions. In addition practitioners still need to undertaken some form of assessment within their specific educational context to ensure that the tutoring that is implemented transfers to their setting. Research on peer tutoring suggested that it has potential to produce consistent positive effects if used in reading with a structured approach, but that studies are not robust enough to ensure that findings transfer and generalise to all contexts. There are also lessons for researchers. As an ‘Academy’ we may not agree what constitutes a peer tutor, a student tutor, a non-professional tutor or a community volunteer. However, if manuscripts define how the authors have interpreted these terms then it is possible to synthesise common research in cognate groups, even if original manuscripts have used differing terms and descriptors initially. There are also methodological issues in design and reporting. Medical RCTs generally follow CONSORT guidelines to ensure consistency of approach and that all appropriate variables are reported (Campbell et al., 2004). There may be a need to develop a strict trial and reporting criteria for educational based RCTs otherwise future reviews will be similarly limited in their ability to provide a definitive evidence base to educational professionals.

References

Agodini, R., & Dynarski, M. (2004). Are experiments the only option? A look at dropout prevention programs. *The Review of Economics and Statisti*cs, *86*(1), 180-194.

Allor, J., & McCathren, R. (2004). The efficacy of an early literacy tutoring program implemented by college students. *Learning Disabilities Research and Practice, 19*(2), 116–129.

Backer, T. (2001). *Finding the balance – program fidelity and adaptation in substance abuse prevention: a state-of-the-art review*. Center for Substance Abuse Prevention, Rockville, MD.

Baker, S., Gersten, R., & Keating, T. (2000). When less may be more: A 2-year longitudinal evaluation of a volunteer tutoring program requiring minimal training. *Reading Research Quarterly, 35*(4), 494–519.

Banerjee, A. V., Cole, S., Duflo, E., & Linden, L. (2007). Remedying education: evidence from two randomized experiments in India. *Quarterly Journal of Economics, 122*(3), 1235- 1264.

Banerjee, A., Banerji, R., Duflo, E., Glennerster, R., & Khemani, S. (2010). Pitfalls of participatory programs: Evidence from a randomized evaluation in education in India. *American Economic Journal: Economic Policy, American Economic Association, 2*(1), 1-30.

Banerjee, A., Banerjee, A. V., & Duflo, E. (2011). *Poor economics: A radical rethinking of the way to fight global poverty*. Public Affairs/Perseus Books Group, New York.

Basch, C. E., Sliepcevich, E. M., Gold, R. S., Duncan, D. F., & Kolbe, L. J. (1985). Avoiding type III errors in health education program evaluations: A case study. *Health Education Quarterly, 12*(4), 315–331.

Beanstalk (2013). *Tips for reading helpers. Retrieved from* www.beanstalkcharity.org.uk/reading-helpers/tips

Becker, B. J., Hedges, L. V., & Pigott, T. D. (2004). *Campbell Collaboration statistical analysis policy brief.* Campbell Collaboration, Oslo, Norway. Retrieved from: www.campbellcollaboration.org/ECG/policy_statasp

Berlin, J. A. (1995). Invited commentary: Benefits of heterogeneity in metaanalysis of data from epidemiologic studies. *American Journal of Epidemiology, 142*, 383-87.

Best Evidence Encyclopedia (2013). *Review methods: Criteria for inclusion in the Best Evidence Encyclopedia*. Retrieved from: www.bestevidence.org/methods/criteria.htm

Bickman, L. (2000). Summing up program theory. *New Directions for Evaluation, 87*, 103-112.

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, *13*, 4-16.

Bold, T., Kimenyi, M., Mwabu, G., Ng'ang'a, A., & Sandefur, J. (2013). *Scaling-up what works: Experimental evidence on external validity in Kenyan education.* Center for Global Development. No. WPS/2013-04. Retrieved from: www.cgdev.org/publication/scaling-what-works

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley, Hoboken, NJ, USA.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed effect and random effects models for meta-analysis. *Research Synthesis Methods*, 1, 97-111.

Boruch, R., May, H., Turner, H., Lavenberg, J., Petrosino, A., De Moya, D., & Foley, E. (2004). Estimating the effects of interventions that are deployed in many places place-randomized trials. *American Behavioral Scientist, 47*(5), 608-633.

Bowling, A. (2009). *Research methods in health: investigating health and health services*. Open University Press, Milton Keynes, UK.

Bray, M. (2006). Private supplementary tutoring: Comparative perspectives on patterns and implications. *Compare 36*(4), 515–30.

Cabana, M., Rand, C., & Powe, N.R. (1999). Why don't physicians follow clinical practice guidelines? A framework for improvement. *Journal of the American Medical Association, 282,* 1458–65.

Cabezas, V., Cuesta, J. I., & Gallego, F. A. (2011). *Effects of short-term tutoring on cognitive and non-cognitive skills: Evidence from a randomized evaluation in Chile.* Poverty Action Lab. Retrieved from: www.povertyactionlab.org/evaluation/impact-short-term-tutoring-cognitive-and-non-cognitive-skills-chile

Campbell, M. K., Elbourne, D. R., & Altman, D. G. (2004). CONSORT statement: extension to cluster randomised trials. *British Medical Journal, 328*(7441), 702-708. Retrieved from: www.bmj.com/content/328/7441/702

Center for Data-Driven Reform in Education (2013). *About the Best Evidence Encyclopedia.* Retrieved from: www.bestevidence.org/aboutbee.htm

Chall, J. S. (1989). Learning to read: The great debate 20 years later. *Phi Delta Kappan, 70,* 521–538.

Chi, M. T. H. (1996). Constructing self-explanations and scaffolded explanations in tutoring. *Applied Cognitive Psychology, 10,* 33-49.

Chi, M. T. H., Bassok, M., Lewis, M., Reimann, P., & Glaser, R. (1989). Self-explanations: how students study and use examples in learning to solve problems. *Cognitive Science, 13,* 145-18.

Chi, M. T. H. Siler, S.A., Jeong, H, Yamauchi, T., & Hausmann, R.G.. (2001). Learning from human tutoring. *Cognitive Science, 25,* 471-533.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, 2nd ed. Laurence Erlbaum Publishers, Hillsdale, New Jersey, USA.

Cohen, P. A., Kulik, J.A., & Kulik C-L.C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal*, *19*, 237-248.

Cooksy, L. J., Gill, P., & Kelly, P. A. (2001). The program logic model as an integrative framework for a multimethod evaluation. *Evaluation and program planning*, *24*(2), 119-128.

Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I., & Petticrew, M. (2008). Developing and evaluating complex interventions: the new Medical Research Council guidance. *British Medical Journal*, *337*, (1655). *Retrieved from:* www.bmj.com/content/337/bmj.a1655.full?maxtoshow=&HITS=10&hits=10&RESULTFORMAT=1&author1=macintyre&andorexacttitle=and&andorexacttitleabs=and&andorexactfulltext=and&searchid=1&FIRSTINDE=

Crévola, C. A., & Hill, P. W. (1998). Evaluation of a whole-school approach to prevention and intervention in early literacy. *Journal of Education for Students Placed at Risk*, *3*(2), 133-157.

Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. O., Hornik, R. C., & Phillips, D. C, (1980). *Toward reform of program evaluation: Aims, methods, and institutional arrangements*. Jossey-Bass, San Francisco, CA..

Dang, H. A., & Rogers F. H. (2008). The growing phenomenon of private tutoring: Does it deepen human capital, widen inequalities, or waste resources? *World Bank Research Observer*, *23*(2), 161-200.

Davis, D. S. (2012). Protocol for a systematic review: Multiple comprehension strategies instruction for improving reading comprehension and strategy outcomes in the middle

grades. *The Campbell Collaboration, Oslo, Norway*. *Retrieved from* www.campbellcollaboration.org/lib/download/2258/

Deeks, J. J., Higgins, J.P.T., & Altman, D.G. (editors). (2011). Chapter 9: Analysing data and undertaking meta-analyses. In: Higgins, J.P.T., Green, S. (editors). *Cochrane handbook for systematic reviews of interventions version 5.1.0.* The Cochrane Collaboration, Oxford, UK.

Delquadri, J., Greenwood, C.R., Whorton, D., Carta, J.J., & Hall, R.V. (1986). Classwide peer tutoring. *Exceptional Children, 52,* 535-542.

Devin-Sheehan, L., Feldman, R.S., & Allen, V. L. (1976). Research on children tutoring children: A critical review. *Review of Educational Research, 46,* 355-385.

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56,* 455-463.

Drugli, M. B., Larsson, B., Fossum, S., & Mørch, W. T. (2010). Five- to six- year outcome and its prediction for children with ODD/CD treated with parent training. *Journal of Child Psychology and Psychiatry, 51*(5), 559-566.

Egger, M., Smith, G.D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315,* 629–634

Elbaum, B., Vaughn, S., Hughes, M.T., & Moody, S.W. (2000). How effective are one-to-one tutoring programs in reading for elementary students at risk for reading failure? A meta-analysis of the intervention research. *Journal of Educational Psychology, 92*(4), 605–619.

Elliott, J., Arthurs, J. & Williams, R. (2000). Volunteer support in the primary classroom: the long-term impact of one initiative upon children's reading performance. *British Educational Research Journal, 26*(2), 227–244.

Field, A. P. (2001). Meta-analysis of correlation coefficients: A Monte Carlo comparison of fixed- and random effects methods. *Psychological Methods*, 6, 161–180.

Fitzgerald, J. (2001) Can minimally trained college student volunteers help young at-risk children to read better? *Reading Research Quarterly, 36,* 28–46.

Fitz-Gibbon, C. (1977). An analysis of the literature of cross-age tutoring. National Institute of Education, Washington, DC. ERIC Document Reproduction Service No. ED 148 807

Flay, B. R. (1986). Efficacy and effectiveness trials (and other phases of research) in the development of health promotion programs. *Preventive medicine, 15*(5), 451-474.

Flay, B. R., Biglan, A., Boruch, R. F., Castro, F.G., Gottfredson, D., Kellam, S., Moscicki, E.K., Schinke, S., Valentine, J.C., & Ji, P. (2005). Standards of evidence: criteria for efficacy, effectiveness and dissemination. *Prevention Science, 6*(3), 151-75.

Flores, J. R. (1989). The effects of cross-age tutoring on underachieving fifth-grade students in the areas of mathematical achievement and self-perception. (Doctoral dissertation). University of Arizona. Retrieved from: arizona.openrepository.com/arizona/handle/10150/184709

Fraser, M. W., Richman, J.M., Galinsky, M.J., & Day, S.H. (2009). *Intervention research: Developing social programs*. Oxford University Press, Oxford, UK.

Fuchs, L. S., Fuchs, D., & Maxwell, L. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education, 9*(2), 20-28.

Fuchs, D., Fuchs, L.S., Thompson, A., Svenson, E., Yen, L., Al Otaiba, S., & Saenz, L. (2001). Peer-assisted learning strategies in reading extensions for kindergarten, first grade, and high school. *Remedial and Special Education*, *22*(1), 15-21.

Gardner, F., Burton, J., & Klimes, I. (2006). Randomised controlled trial of a parenting intervention in the voluntary sector for reducing child conduct problems: Outcomes and mechanisms of change. *Journal of Child Psychology & Psychiatry, 47*, 1123-1132.

Gardner, F., Hutchings, J., Bywater, T., & Whitaker, C. (2010). Who benefits and how does it work? Moderators and mediators of outcomes in a randomised trial of parenting interventions in multiple 'Sure Start' services. *Journal of Clinical Child & Adolescent Psychology, 39*, 568-580.

Gersten, R., Baker, S., & Lloyd, J. W. (2000). Designing high-quality research in special education group experimental design. *The Journal of Special Education*, *34*(1), 2-18.

Ginsburg-Block, M. D., Rohrbeck, C.A., & Fantuzzo, J.W. (2006). A meta-analytic review of social, self-concept, and behavioral outcomes of peer-assisted learning. *Journal of Educational Psychology*, *98*(4), 732–749.

Glazerman, S., Levy, D.M., & Myers, D. (2003). Noexperimental versus experimental estimates of earnings impacts. *Annals of the American Academy of Political and Social Science*, *589*(1), 63-93.

Goodlad, S., & Hirst, B. (1989) *Peer tutoring: A guide to learning by teaching*. Kogan Page, London..

Goodlad, S. (1995). *Students as tutors and mentors*. Kogan Page, London.

Goodlad, S. (1998). *Mentoring and tutoring by students*. Kogan Page, London.

Gøtzsche, P. C., Hr´objartsson, A., Maric, K., & Tendal, B. (2007). Data extraction errors in meta-analyses that use standardized mean differences. *Journal of the American Medical Association*, *298*, 430–437.

Graesser, A. C. (2009). Inaugural editorial for Journal of Educational Psychology. *Journal of Educational Psychology*, *101* (2), 259–261.

Graesser, A. C., D'Mello, S., & Cade, W. (2009). Instruction based on tutoring. In Mayer, R.E. & Alexander, P.A. (Eds.), *Handbook of Research on Learning and Instruction*. Routledge Press, New York

Greenwood, C. (1997). Classwide peer tutoring. *Behavior and Social Issues, 7*, 53-58.

Guyatt G. H., Haynes R.B., Jaeschke R.Z., Cook D.J., Green L., Naylor C.D., Wilson M.C., & Richardson W.S. (2000). Users guide to the medical literature XXV. Evidence-based medicine: Principles for applying the users guides to patient care. *Journal of the American Medical Association, 284*, 1290–1296.

Guyatt, G. H., Oxman, A.D., Kunz, R., Falck-Ytter, Y., Vist, G.E., Liberati, A., & Schünemann, H.J., GRADE Working Group. (2008a). Rating quality of evidence and strength of recommendations: Going from evidence to recommendations. *British Medical Journal, 336*(7652), 1049-1051.

Guyatt, G. H., Oxman, A.D., Kunz, R., Jaeschke, R., Helfand, M., Liberati, A., Vist GE, Schünemann, H.J., GRADE working group. (2008b). Rating quality of evidence and strength of recommendations: Incorporating considerations of resources use into grading recommendations. *British Medical Journal, 336*(7654), 1170-1173.

Ham, W. (1977). *Effects of a volunteer tutor program on self-esteem and basic skills achievement: In the primary grades of a southern rural school system.* (Unpublished . dissertation). University of Florida.

Hargreaves, D. (1996). Teaching as a research-based profession: Possibilities and prospects. The teacher training agency annual lecture 1996. Retrieved from: eppi.ioe.ac.uk/cms/Portals/0/PDF%20reviews%20and%20summaries/TTA%20Hargreaves%20lecture.pdf

Hartley, S. S. (1977). Meta-analysis of the effects of individually paced instruction in mathematics. (Doctoral dissertation). University of Colorado. *Dissertation Abstracts International*, 38(7-A), 4003, University Microfilms (77-29).

Hawe, P., Shiel, A., & Riley, T. (2004). Complex interventions: How "out of control" can a randomized controlled trial be? *British Medical Journal, 328*, 1561-1563.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group randomized trials in education. *Educational Evaluation and Policy Analysis*, 29, 60-87.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press, Orlando, FL, USA.

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, *327*(7414), 557-60.

Higgins J. P., Altman D.G. (2008). Chapter 8: Assessing risk of bias in included studies. In Higgins, J.P.T., & Green, S. (Eds.), *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons, Chichester, UK..

Higgins, J. P. T., & Green, S. (2011). *Cochrane handbook for systematic reviews of interventions (Version 5.1.0).* Retrieved from: handbook.cochrane.org/

Higgins, S., Katsipataki, M., Kokotsaki, D., Coleman, R., Major, L.E., & Coe, R. (2013). *The Sutton Trust-Education Endowment Foundation teaching and learning toolkit*. Education Endowment Foundation, London .Retrieved from*:* educationendowmentfoundation.org.uk/toolkit

Hopewell, S., Loudon, K., Clarke, M.J., Oxman, A.D., Dickersin, K. (2009). Publication bias in clinical trials due to statistical significance or direction of trial results. *Cochrane Database of Systematic Reviews*, 1 (MR000006). DOI:10.1002/14651858.MR000006.pub3.

Ioannidis, J. P., Patsopoulos, N. A., & Evangelou, E. (2007). Uncertainty in heterogeneity estimates in meta-analyses. *British Medical Journal, 335*(7626), 914-916.

Ireson, J. (2004). Private tutoring: How prevalent and effective is it? *London Review of Education, 2*(2), 109-122.

Iyenger, S., & Greenhouse, J.B. (1988). Selection models and the file drawer problem (with discussion). *Statistical Science, 3,* 109-135.

Jensen, R. J. (1991). The effects of cross-age tutoring on the reading achievement of underachieving second and fifth-grade students. (Doctoral dissertation). Brigham Young University. *ProQuest Dissertations and Theses*, 208-208. Retrieved from*:* search.proquest.com/docview/303975464?accountid=13042

Kjaergard, L. L., Villumsen, J., & Gluud, C. (2001). Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. *Annals of Internal Medicine, 135*, 982–989.

Koh, S., Sanders, K,, & Meyer, J. (2012). Roles of active learning and tutor input in students' perception of learning. *Teaching and Learning Forum*, 1-9. Retrieved from: www.roger-atkinson.id.au/tlf2012/refereed/koh.pdf

Krishnaratne, S., White H., & Carpenter, E. (2013). Quality education for all children? What works in education in developing countries. *3ie*. Retrieved from*:* www.3ieimpact.org/en/evaluation/working-papers/working-paper-20/

Lally, P., van Jaarsveld, C. H. M., Potts, H. W. W. and Wardle, J. (2010). How are habits formed: Modelling habit formation in the real world. *European Journal of Social Psychology, 40*: 998–1009.

Lau, J., Ioannidis, J. P., & Schmid, C. H. (1998). Summing up evidence: one answer is not always enough. *Lancet, 351* (9096), 123-127.

Learning Together (2013). Reading together elementary. *Retrieved from:* www.learningtogether.com/reading/elementary_reading.html

Lee, C. C. (1980). The homework helper program: Volunteer service for academic and social enrichment in the elementary school, *The School Counselor*, *28*, 11–21.

Lee, Y. S., Morrow-Howell, N., Jonson-Reid, M., & McCrary, S. (2012). The effect of the experience corps[R] program on student reading outcomes. *Education and Urban Society*, *44*(1), 97-118.

LeLorier, J., Gregoire, G., Benhaddad, A., Lapierre, J., & Derderian, F. (1997). Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *New England Journal of Medicine*, *337*(8), 536-542.

Leung, K. C., Marsh, H. W., & Craven, R. G. (2005). Are peer tutoring programs effective in promoting academic achievement and self-concept in educational settings: A meta-analytical review. International Conference of the Australian Association for Research in Education, Sydney, Australia.

Li, T., Han, L., Rozelle, S., & Zhang, L. (2010). Cash incentives, peer tutoring, and parental involvement: A study of three educational inputs in a randomized field experiment in China. *Peking University*, Beijing, China. Retrieved from: mitsloan.mit.edu/neudc/papers/paper_223.pdf

Linnan, L., & Steckler, A. (2002). Process evaluation for public health interventions and research: An Overview. In Steckler, A, & Linnan, L. (Eds.) Process evaluation for public health interventions. Jossey-Bass, San Francisco, CA.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical Meta-Analysis*. Sage, Thousand Oaks, CA.

Lipsey, M. W. (2005). The challenges of interpreting research for use by practitioners: comments on the latest products from the Task Force on Community Preventive Services. *American Journal of Preventative Medicine 28*, 1– 3.

Littell, J. H., Corcoran, J.C., & Pillai, V. (2008). Systematic reviews and meta-analysis. Oxford University Press, Oxford, UK.

Loyalka, P., Liu, C., Song, Y., Yi, H., Huang, X., Wei, J., Zhang, L, Shi, Y., Chu, J., & Rozelle, S. (2013). Can information and counseling help students from poor rural areas go to high school? Evidence from China. *Journal of Comparative Economics*, *36*, 26-40.

Loenen, A. (1989). The effectiveness of volunteer reading help and the nature of the reading help provided in practice. *British Educational Research Journal*, *15*, 297–316.

Lomas, J. (1990). Finding audiences, changing beliefs: the structure of research use in Canadian health policy. *Journal of Health Politics, Policy and Law*, *15*(3), 525-542.

Margolis, H. (2005). Increasing struggling learners' self-efficacy: what tutors can do and say. *Mentoring and Tutoring: Partnership in Learning*, *13*(2), 221–238.

Marlowe, D. B., Festinger, D. S., Arabia, P. L., Dugosh, K. L., Benasutti, K. M., & Croft, J. R. (2009). Adaptive interventions may optimize outcomes in drug courts: A pilot study. *Current Psychiatry Reports*, *11*(5), 370-376.

McEwan, P. J. (2012). Cost-effectiveness analysis of education and health interventions in developing countries. *Journal of Development Effectiveness*, *4*(2), 189-213.

McEwan, P. J. (2013). Improving learning in primary schools of developing countries: A meta-analysis of randomized experiments. Retrieved from: academics.wellesley.edu/Economics/mcewan/PDF/meta.pdf

Medway, F. J. (1995). Tutoring. In L.W. Anderson (Ed.), *International Encyclopedia of Teaching and Teacher Education*, Pergamon, Cambridge.

Meier, J. D., & Invernizzi, M. (2001). Book buddies in the Bronx: Testing a model for America Reads. *Journal of Education for Students Placed at Risk*, *6*(4), 319-333.

Merrill, D. C., Reiser, B. J., Merrill, S. K., & Landes, S. (1995). Tutoring: guided learning by doing. *Cognition and Instruction, 13*(3), 315–372.

Mihalic, S. (2004). The importance of implementation fidelity. *Emotional and Behavioral Disorders in Youth, 4*(83-86), 99-105.

Miller, D., Topping, K., & Thurston, A. (2010). Peer tutoring in reading: The effects of role and organization on two dimensions of self- esteem. *British Journal of Educational Psychology, 80*(3), 417-433.

Miller, S., Connolly, P., Odena, O., & Styles, B. (2009). A randomised controlled trial evaluation of business in the community's time to read pupil mentoring programme. *Centre for Effective Education, Queen's University Belfast.* Retrieved from: www.qub.ac.uk/cee

Miller, S., & Connolly, P. (2012). A randomised controlled trial evaluation of time to read, a volunteer tutoring program for 8- to 9-year-olds. *Educational Evaluation and Policy Analysis, 35*(1), 23-37.

Miller, S., Connolly, P., & Maguire, L.K. (2012). The effects of a volunteer mentoring programme on reading outcomes among eight- to nine-year-old children: A follow up randomized controlled trial. *Journal of Early Childhood Research, 10,* 134-144.

Miller, S., Maguire, L. K., & Macdonald, G. (2012). Home-based child development interventions for preschool children from socially disadvantaged families. *Campbell Systematic Reviews, 1.* DOI: 10.4073/csr.2012.1

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D.G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med 6*(7): e1000097. Retrieved from: www.plosmedicine.org/article/info%3Adoi%2F10.1371%2Fjournal.pmed.1000097. DOI:10.1371/journal.pmed.1000097

Mullen, E. J. (2006). Choosing outcome measures in systematic reviews: Critical challenges. *Research on Social Work Practice, 16*(1), 84-90.

Newman, M. (2003). A pilot systematic review and meta-analysis on the effectiveness of problem based learning. *Campbell Collaboration Systematic Review Group on the effectiveness of problem based learning*. Learning and Teaching Support Network, University of Newcastle, Newcastle, UK.

Oakley, A., Strange, V., Bonell, C., Allen, E., & Stephenson, J. (2006). Health services research: process evaluation in randomised controlled trials of complex interventions. *British Medical Journal, 332*(7538), 413-416.

O'Connor, D., Green, S., Higgins, J.P. (2008). Chapter 5: Defining the review question and developing criteria for including studies. In Higgins, J.P.T., & Green, S. (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions*. John Wiley & Sons, Chichester, UK.

Otaiba, S. A., & Schatschneider, C. (2005). Tutor-assisted intensive learning strategies in kindergarten: How much is enough? *Exceptionality, 13*(4), 195-208.

Palincsar, A. S., & Brown, A. L. (1989). Classroom dialogues to promote self-regulated comprehension. Brophy, J. (Ed.), *Teaching for meaningful understanding and self-regulated learning*, 1. JAI Press, Greenwich, CT, USA.

Pesci, A. (2015). *Cooperative learning and peer tutoring to promote students' mathematics education*. Retrieved from*:* math.unipa.it/~grim/21_project/Pesci486-490.pdf

Petrosino, A., Morgan, C., Fronius, T.A., Tanner-Smith, E.E., & Boruch, R.F. (2012). Interventions in developing nations for improving primary and secondary school enrolment of children: A systematic review. Campbell Systematic Reviews 2012:19. DOI: 10.4073/csr.2012.19

Petticrew, M., & Roberts, H. (2003). Evidence, hierarchies, and typologies: horses for courses. *Journal of Epidemiology and Community Health, 57*(7), 527-9.

Pigott, T. (2011). Cluster adjustments in computing effect sizes. *Center for Evidence-Based Crime Policy-Campbell collaboration joint symposium on evidence-based policy, George Mason University*. Retrieved from: www.campbellcollaboration.org/resources/training/advanced_methods.php

Pinnell, G. S., Lyons, C. A., Deford, D. E., Bryk, A. S., & Seltzer, M. (1994). Comparing instructional models for the literacy education of high-risk first graders. *Reading Research Quarterly, 29*, 8-39

Policy Studies Associates (2007). Evidence of long-term learning outcomes among reading together tutees. Washington, DC

Poole, C., & Greenland, S. (1999). Random-effects meta-analyses are not always conservative. *American Journal of Epidemiology, 150*(5), 469-475.

Potter, J. (1994). 'No Limit' a blueprint for involving volunteer tutors in primary schools. *Mentoring & Tutoring: Partnership in Learning, 2*(2), 61-62.

Poverty Action Lab (2009). Read India: Helping primary school students in India acquire basic reading and math skills. Retrieved from: www.povertyactionlab.org/evaluation/read-india-helping-primary-school-students-india-acquire-basic-reading-and-math-skills

Pridmore, P., Stephens, D., & Stephens, J. (2000). *Children as partners for health: a critical review of the child-to-child approach*. Zed Books, London.

Pritchard, D. (1976). the effect of cross-age tutoring on adolescence: An inquiry into theoretical assumptions. *Review of Educational Research, 46* (2), 215-237.

Promise Neighborhoods Research Consortium (2010a). *Volunteer tutoring programs.* Retrieved from: promiseneighborhoods.org/policies/volunteer-tutoring-programs/

Promise Neighborhoods Research Consortium (2010b). *Peer-to-peer tutoring.* Retrieved from*:* promiseneighborhoods.org/kernels/peer-peer-tutoring/

Pullen, P. C., Lane, H. B., & Monaghan, M. C. (2004). Effects of a volunteer tutoring model on the early literacy development of struggling first-grade students. *Reading Research and Instruction*, *43*(4), 21–40.

Reisner, E., Petry, C., & Armitage, M. (1989). *A review of programs involving college students as tutors or mentors in grades k-12.* Policy Studies Associates, Inc., Department of Education, Washington, DC.

Rimm-Kaufman, S. E., Kagan, J., & Byers, H. (1998). The effectiveness of adult volunteer tutoring on reading among "at risk" first-grade children. *Reading Research and Instruction*, *38*(2), 143–152.

Ritter, G. W. (2000). The academic impact of volunteer tutoring in urban public elementary schools: Results of an experimental design evaluation. (Doctoral dissertation). University of Pennsylvania. Retrieved from: *Dissertation Abstracts International*, 61(3A).

Ritter, G. W., & Maynard, R. A. (2008). Using the right design to get the "wrong" answer? Results of a random assignment evaluation of a volunteer tutoring program. *Journal of Children's Services*, *3*(2), 4–16.

Ritter, G. W., Barnett, J. H., Denny, G. S., & Albin, G. R. (2009). The effectiveness of volunteer tutoring programs for elementary and middle school students: A meta-analysis. *Review of Educational Research*, *79*(1), 3–38.

Ritter, G. W., Denny, G., Albin, G., Barnett, J., & Blankenship, B. (2006). The effectiveness of volunteer tutoring programs: a systematic review. *Campbell Collaboration*, 7. doi: 10.4073/csr.2006.7

Robinson, D. R., Schofield, J. W., & Steers-Wentzell, K. L. (2005). Peer and cross-age tutoring in math: outcomes and their design implications. *Educational Psychology Review*, *17*(4), 327–362.

Rohrbeck, C. A., Ginsburg-Block, M. D., Fantuzzo, J. W., & Miller, T. R. (2003). Peer-assisted learning interventions with elementary school students: A meta-analytic review. *Journal of Educational Psychology*, *95*, 240–257.

Roscoe, R. D., & Chi, M. T. (2007). Understanding tutor learning: Knowledge-building and knowledge-telling in peer tutors' explanations and questions. *Review of Educational Research*, *77*(4), 534-574.

Rosenshine, B., & Furst, N. (1969). The effects of tutoring upon pupil achievement: A research review. Washington, D.C.: Office of Education. *ERIC Document Reproduction Service*, ED 064462.

Rychetnik, L., Frommer, M., Hawe, P., & Shiell, A. (2002). Criteria for evaluating evidence on public health interventions. *Journal of Epidemiology and Community Health*, *56*(2), 119-127.

Shadish, W. R., Cook, T.D., & Campbell, D.T. (2002*). Experimental and quasi-experimental designs for generalized causal inferenc*e. Houghton-Mifflin , Boston, MA..

Shanahan, T., & Barr, R. (1995). Reading Recovery: An independent evaluation of the effects of an early instructional intervention for at-risk learners. *Reading Research Quarterly*, *30*, 958-997.

Shanahan, T. (1998). On the effectiveness and limitations of tutoring in reading. *Review of Research in Education, 23,* 217-234.

Slavin, R. E., & Smith, D. (2009). The relationship between sample sizes and effect sizes in systematic reviews in education. *Educational Evaluation and Policy Analysis*, *31*(4), 500-506.

Slavin, R. E., & Lake, C. (2008a). Effective programs in elementary mathematics: A best-evidence synthesis. *Review of Educational Research, 78* (3), 427-515.

Slavin, R. E. (2008b). Perspectives on evidence-based research in education—What works? Issues in synthesizing educational program evaluations. *Educational Researcher, 37*(1), 5-14.

Slavin, R.E., Lake, C., Chambers, B., Cheung, A., & Davis, S. (2009a). *Effective beginning reading programs.* Center for Data-Driven Reform in Education, Johns Hopkins University, Baltimore, MD, USA..

Slavin, R.E., Lake, C., Cheung, A., & Davis, S. (2009b). *Beyond the basics: Effective reading programs for the upper elementary grades.* Center for Data-Driven Reform in Education, Johns Hopkins University, Baltimore, MD, USA.

Slavin, R.E., Lake, C., Chambers, B., Cheung, A., & Davis, S. (2009c). Effective reading programs for the elementary grades: A best-evidence synthesis. *Review of Educational Research, 79* (4), 1391-1466.

Slavin, R.E., Lake, C., Davis, S., & Madden, N. (2010). Effective programs for struggling readers: A best evidence synthesis. *Educational Research Review*. Retrieved from*:* dx.doi.org/10.1016/j.edurev.2010.07.002

Slavin, R.E., & Madden, N.A. (2011). Measures inherent to treatments in program effectiveness reviews. *Journal of Research on Educational Effectiveness, 4* (4), 370-380.

Slavin, R.E., Lake, C, Davis, S., & Madden, N.A. (2011). Effective programs for struggling readers: A best-evidence synthesis. *Educational Research Review, 6*(1), 1-26.

Society for Prevention Research, Standards of Evidence Committee (2004*). Standards of evidence: Criteria for efficacy, effectiveness and dissemination*. Retrieved from: www.preventionresearch.org/StandardsofEvidencebook.pdf

Spörer, N., Brunstein, J.C., & Kieschke, U. (2009). Improving students' reading comprehension skills: Effects of comprehension instruction and reciprocal teaching. *Learning and Instruction, 19*, 272-286.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *21*, 361–407.

Sterne, J. A., Egger, M., & Smith, G. D. (2001). Investigating and dealing with publication and other biases in meta-analysis. *British Medical Journal*, *323*(7304), 101-105.

Sutton, A. J., Duval, S. J., Tweedie, R. L., Abrams, K.R., & Jones, D. R. (2000a). Empirical assessment of effect of publication bias on meta-analyses. *British Medical Journal, 320,* 1574-1577.

Sutton, A. J., Duval, S. J., Tweedie, R. L., Abrams, K.R., & Jones, D. R. (2000b). High false positive rate for trim and method. *British Medical Journal*. *320,* 1574-1577. Retrieved from: www.bmj.com/rapid-response/2011/10/28/high-false-positive-rate-trim-and-fill-method

Swanson, H. L., & Hoskyn, M. (1998). Experimental intervention research on students with learning disabilities: A meta-analysis of treatment outcomes. *Review of Educational Research*, *68*(3), 277-321.

Thorlund, K., Imberger, G., Johnston, B. C., Walsh, M., Awad, T., Thabane, L., Gluud, C., Devereaux, P. J., & Wetterslev, J. (2012). Evolution of heterogeneity (I^2) estimates and their 95% confidence intervals in large meta-analyses. *PLoS One* 7(7), e39471

Topping, K. J. (1998). Commentary: Effective tutoring in America Reads: A reply to Wasik. *The Reading Teacher , 52*(1), 42-50.

Topping, K. J. (2000). Tutoring. Educational practices series, 5. *International Bureau of Education*. Retrieved from*:* une*SD*oc.unesco.org/images/0012/001254/125454e.pdf

Topping, K. J. (2004) Tutoring in mathematics: a generic method. *Mentoring and Tutoring: Partnership in Learning, 12*(3), 353-370.

Topping, K., & Whiteley, M. (1993). Sex differences in the effectiveness of peer tutoring. *School Psychology International, 14*(1), 57-67.

Topping, K. J., & Hill, S. (1995). University and college students as tutors for schoolchildren: A typology and review of evaluation research, 13-31. In Goodlad, S. (Ed.) *Students as tutors and mentors*. Kogan Page, London.

Topping, K. J., Miller, D., Murray P., & Conlin, N. (2011). Implementation integrity in peer tutoring of mathematics. *Educational Psychology: An International Journal of Experimental Educational Psychology, 31*(5), 575-593.

Topping, K. J., Thurston, A., McGavock, K., & Conlin, N. (2012). Outcomes and process in reading tutoring. *Educational Research, 54*(3), 239-258.

Torgerson, C. J., Torgerson, D. J., Birks, Y. F., & Porthouse, J. (2005). A comparison of randomised controlled trials in health and education. *British Educational Research Journal*, *31*(6), 761-785.

Torgerson, C. J. (2006). The quality of systematic reviews of effectiveness in literacy learning in English: a 'tertiary' review. *Journal of Research in Reading*, *30*(3), 287-315.

Torgerson, C. J., & King, S. (2002). Do volunteers in schools help children learn to read? A systematic review of randomised controlled trials. *Educational Studies*, *28*(4), 433-444. Retrieved from: www.tandfonline.com/doi/abs/10.1080/0305569022000042435

Tymms, P., Merrell, C., Thurston, A., Andor, J., Topping K., & Miller, D. (2011). Improving attainment across a whole district: school reform through peer tutoring in a randomized controlled trial, *School Effectiveness and School Improvement*, *22*(3), 265-289

Vadasy, P.F., Jenkins, J.R., Antil, L.R., Phillips, N.B., & Pool, K. (1997). The Research-to-practice ball game. Classwide peer tutoring and teacher interest, implementation, and modifications. *Remedial and Special Education*, *18*(3), 143-156.

Vygotsky, L.S. (1978). *Mind in Society*. Harvard University Press, Cambridge, MA.

Walker, R., Hoggart, L., & Hamilton, G. (2008). Observing the implementation of a social experiment. *Evidence & Policy: A Journal of Research, Debate and Practice*, *4*(3), 183-203.

Wasik, B. A., & Slavin, R.E. (1993). Preventing early reading failure with one-to-one tutoring: A review of five programs. *Reading Research Quarterly*, *28*, 179-200

Wasik, B. A. (1998). Volunteer tutoring programs in reading: A review. *Reading Research Quarterly*, *33*(3), 266-291.

Webb, N. M. (1989). Peer interaction and learning in small groups. *International Journal of Education Research*, *13*, 21-39

Weiss, C. H. (1997). Theory- based evaluation: Past, present, and future. *New Directions for Evaluation*, *76*, 41-55

What Works Clearing House (2010). Procedures and standards handbook (Version 2.1). What Works Clearing House, Washington, DC..

Wilde, E. T., & Hollister, R. (2007). How close is close enough? Evaluating propensity score matching using data from a class size reduction experiment. *Journal of Policy Analysis & Management*, *26*(3), 455-477.

Wilson, D., & Lipsey, M. (2001). The role of method in treatment effectiveness research: Evidence from meta-analysis. *Psychological Methods*, *6*(4), 413-429

Zief, S. G., Lauver, S., & Maynard, R.A. (2006) *Impacts of after-school programs on student outcomes.* Campbell Library of Systematic Reviews, Campbell Collaboration, Oslo, Norway.

Zucker, D. M., Lakatos, E., Webber, L. S., Murray, D. M., McKinlay, S. M., Feldman, H. A., Kelder, S.H., & Nader, P. R. (1995). Statistical design of the child and adolescent trial for cardiovascular health (CATCH): implications of cluster randomization. *Controlled Clinical Trials, 16*(2), 96-118.

Table 1. Overview of Key Features of the Included Studies

Authors of study	*N*	Description of tutees	Description of tutors	Total	Per week	Length (weeks)	Fidelity	Location	Intervention description
				in hours					
Allor & McCathren (2004)	86 *year 1*	Gr.1 *M*= 6.7 y.o.	University education major student volunteers	12	1	26	Used a checklist *M*= 86.98 % (*SD* = 5.67)	8 underachieving schools, urban south USA	- Outside class during school day - Remedial tutoring for low-achieving children - Tutor training: America Reads tutor training, 3 1-hour trainings, monthly training, and on-site assistance - Scripted lessons with progressively challenging lessons, containing games on phonemic awareness, letter-sound correspondence, word-study activities and reading of levelled books - 3 research assistants observed and supported tutors
Allor & McCathren (2004)	157 *year 2*	Gr.1 *M*= 6.6 y.o.		13	1	26	*M*= 86.53% (*SD* = 4.80)	10 underachieving schools	
Baker, Gersten & Keating (2000)	84	Gr.1	Adult community volunteers (33% 30-45 y.o., 29% 45-65, 20% > 65)	37	1	72	Not reported	6 Title-1 schools, Oregon, USA	- Outside class during school day - Remedial tutoring for low-achieving children - Tutees selected based on reading difficulties and need for relationship with a caring adult - Tutor training: 1-2 hour training and community volunteer handbook - Tutoring focused on increasing children's interest in reading, program providing books for children to take home.
Cabezas, Cuesta & Gallego 2011	4903	Gr.4 9-10 y.o.	University student volunteers	18	1.5	12	High volunteer turnover	85 vulnerable schools in 10 counties in Biobio and Great Santiago regions, Chile	- After class - School-wide one to small group tutoring (5-6 students assigned to a tutor) - Tutoring focused on "shared-reading … of traditional stories and informative texts, which are age-and interest appropriate for students" - Volunteers supported by an employee of "Fundación para la Superación de la Pobreza" at each school - Volunteers received stipends for travel

Authors of study	*N*	Description of tutees	Description of tutors	Total	Per week	Length (weeks)	Fidelity	Location	Intervention description
				in hours					
Elliott, Arthurs & Williams (2000)	30	Reception class 4-5 y.o.	Adult community volunteers	19	1	19	Didn't measure	3 low-SES schools, Northeast England, UK	"Time for Reading" - During school day, both in and outside classroom - Class-wide tutoring one to small group tutoring - Tutor training: 6 hours over 3 weeks - Tutors worked alongside classroom teacher, providing "individual assistance … The focus of the work was reading for meaning and most of the training sessions involved the child reading to the helper from a fiction text and discussing elements of the story"
Ham (1977)	147	Gr. 1, 2, 3	Adult community volunteers,	36	2	22	Record keeping failed, high tutor turnover	4 schools with low SES & minority students, Sumter County, rural USA	- During language arts classes, outside class - One-to-one and small groups tutoring - Remedial tutoring for low-achieving children - Tutors worked following teachers' recommendations, "because of the turnover in volunteers and because volunteers as persons are difficult to program or control, plans for standardization of instructional approach had to be abandoned" p. 63
Jensen (1991)	93	Gr.2	Gr. 5	46	2	23	Not reported	7 elementary schools, Cache Valley, Utah, USA	- One-to-one tutoring - Remedial tutoring for low-achieving children - Tutor training: weekly sessions on "effective tutoring techniques, error correction procedures, and proper prompting techniques"; effects on tutors also assessed -Tutoring focused on timed reading aloud, reading passages assigned by paraprofessionals; tutors corrected mistakes and feedback for correct reading, asked comprehension questions
Lee (1980)	40	Gr.3-6	University volunteers, juniors and seniors	76	4	19	Not reported	4 schools, low SES & minority, urban USA	- After class - One to small group tutoring - Remedial tutoring for low-achieving children, or based on minority status or residence

Authors of study	*N*	Description of tutees	Description of tutors	Total	Per week	Length (weeks)	Fidelity	Location	Intervention description
				in hours					
									- Tutoring focused on homework assignments, improving reading and maths skills, addressing personal concerns - Tutor training: 7 training modules; tutors supervised by two graduate counselling students
Lee Morrow-Howell, Jonson-Reid & McCrary (2012)	881	Gr.1, 2, 3 *M*=7.09 y.o.	Adult community volunteers, 50 to 93 y.o., mean 65	21	1.75	36	Not reported	81 schools in Boston, 52 in New York, and 41 in Port Arthur, USA	"Experience Corps" - One-to-one tutoring - Remedial tutoring for low-achieving children - Tutor training: 15 to 32 hours - NY: Book Buddies (phonics, rereading familiar books, word study, writing, and reading a new book) - Boston: Reading Coaches (building student's oral vocabulary and increasing reading comprehension by asking prediction questions, discussing, and writing about the story) - Port Arthur: Brigance Inventory of Basic Skills materials (word recognition, comprehension, and word analysis) Nationally, 43% of community volunteers have high school diplomas, and 75% –some college education, some are former teachers
Loenen (1989)	81	7-11 y.o., *M*= 8.8 y.o.	Adult community volunteers	24	1	26	Observed 15 tutors, low fidelity to the training	13 schools in inner London, UK	"Volunteer Reading Help" - Outside class during school day - One-to-one tutoring - Remedial tutoring for low-achieving children - Tutor training: short compulsory training course (3 1.5- sessions on reading & practical tips) - Volunteers encouraged to talk to teachers, but no formal structure

Authors of study	*N*	Description of tutees	Description of tutors	Total	Per week	Length (weeks)	Fidelity	Location	Intervention description
				in hours					
Miller, Connolly, Odena & Styles (2009)	734	8-9 y.o.	Adult community volunteers	13	0.5	58	High tutor turnover, "variation in delivery"	Northern Ireland, UK 50 schools	"Time to Read" - Outside class during school day - One-to-one tutoring - Remedial tutoring for below-average performing children - Tutor training: half-day tutor training in paired reading strategies to improve reading fluency, word recognition, meaning, and comprehension for tutors, emphasizing repetition, alternate reading, word recognition, word meaning and comprehension, no structure provided for the sessions but a set of books. Some children received a workplace visit.
Miller,Connolly & Maguire (2012)	483	8-9 y.o.	Adult community volunteers	29	1	29	Not recorded	50 schools in Northern Ireland	"Time to Read" See above (note increased intensity/dose)
Policy Studies Associates (2007)	124	Gr.2	Gr. 4-5	72	2	36	Not recorded	Irving, TX, and Mongomery County, Maryland, US	"Reading Together" - Outside class during school day - One-to-one tutoring - Remedial tutoring for students at risk of reading failure, - Tutor training: 9 hours - Tutoring focused on a curriculum on "reading comprehension, reading fluency, vocabulary, and writing ... to move students from decoding to comprehending"
Pullen, Lane & Monaghan (2004)	47	Gr.1	University student volunteers, majors related to education	10	0.75	12	Used a checklist *M*=92%	North-central Florida, US 10 schools	- Outside class during school day - One-to-one tutoring - Remedial tutoring for students below 30th percentile - Tutor training: 4 hours - Three-step tutoring model: repeated reading of familiar text, explicit coaching in decoding and word-solving strategies, and reading new books

Authors of study	*N*	Description of tutees	Description of tutors	Total	Per week	Length (weeks)	Fidelity	Location	Intervention description
				in hours					
									during each session
Rimm-Kaufman, Kagan & Byers, (1998)	42	Gr.1	Community volunteers	72	2.25	35	Not reported	Cambridge, MA, US 6 schools	- Outside class during school day - One-to-one tutoring - Remedial tutoring for students below 30th percentile - Tutor training: 5 sessions and bimonthly meetings - Prescribed tutoring session schedule: reading for meaning associations between print and pictures, phonetics taught within the context of stories). "The tutors used games, drawing, writing, and related activities to engage the children in learning".
Ritter (2000)	319	At-risk Gr.2, 3,4, 5	University volunteers	21	1	21	Not reported	Philadelphia, PA, US 11 schools	"West Philadelphia Tutoring Project" - Outside class during school day - One-to-one tutoring - Remedial tutoring for students below 30th percentile - Tutor training: minimal training and supervision - Limited tutoring structure - "variety of tasks … spelling, reading, math problems, games, puzzles, crafts, and storytelling"

*SES-socioeconomic status, y.o.- years old

Table 2. Effect Sizes and random effects of included studies

Outcome area	*N* cohorts	*N* students	Hedges' *g* (random effects)	95% CI	*p*-value	Heterogeneity
Composite measure of reading	16	8251	**0.18***	0.08, 0.27	<0.001	Q=97.8; df=15; p=0.000; I^2=84.663; T=0.155; T^2=0.024
Overall reading ability measure	6	1457	0.07	-0.06, 0.20	0.299	Q=7.903; df=5; p=0.162; I^2=36.737; T=0.095; T^2=0.009
Decoding measure	9	7081	**0.29***	0.13, 0.44	0.000	Q=60.095; df=8; p=0.000; I^2=86.688; T=0.208; T^2=0.043
Comprehension measure	10	6945	**0.11***	0.01, 0.21	0.025	Q=15.223; df=9; p=0.085; I^2=40.877; T=0.091; T^2=0.008
Fluency measure	4	687	0.11	-0.21, 0.44	0.494	Q=13.104; df=3; p=0.004; I^2=77.106; T=0.275; T^2=0.075
Writing measure	3	4975	0.01	-0.07, 0.09	0.774	Q=0.281; df=2; p=0.869; I^2=0.000; T=0.000; T^2=0.000
Mathematics measure	3	506	-0.02	-0.18, 0.13	0.778	Q=1.774; df=2; p=0.412; I^2=0.000; T=0.000; T^2=0.000

** Significantly different from zero, p < .05, favouring tutoring over the control.*

Table 3 Reading Effect Sizes by moderator

Study feature	*N* cohorts	Hedges' *g* (random effects)	95% CI	Homogeneity between groups (random effects)
Study size				
Large	5	0.08	0.04, 0.11	Q=9.771*, df=2, p=0.008
Small	11	0.23	0.07, 0.39	
Publication status				
Journal article	11	0.21	0.08, 0.34	Q=0.619, df=0.6, p=0.431
Report or dissertation	5	0.13	-0.03, 0.28	
Type of tutor				
Older child peer tutor	2	0.24	-0.07, 0.55	
University student	6	0.28	0.03, 0.53	Q=2.230, df=2, p=0.328
Adult community volunteer	8	0.11	0.03, 0.18	
Tutoring structure				
Loosely structured	9	0.33	0.14, 0.52	Q=5.903*, df=1, p=0.015
Highly structured	7	0.08	-0.01, 0.16	

Table 4. Non-academic outcomes in the included studies

Study	Outcome	Scale	Hedges *g* (95% CI)
Lee 1980	Self-concept of reading	Piers-Harris Children's Self-Concept Scale	0.31 (-0.32, 0.93)
	Classroom behaviour	Devereaux Elementary School Behavior Rating Scale	-2.12 (-2.90, -1.35)
Loenen 1989	General self-concept	McDaniel-Piers Young Children's Self-concept Scale	0.06 (-0.39, 0.51)
	Composite classroom behavior	Rutter B-scale for teachers	-0.10 (-0.58, 0.39)
Miller 2009	Future aspirations	Future aspirations (Loeber, Stouthamer-Loeber, Van Kammen & Farrington, 1991))	0.17* (0.02, 0.33)
	Enjoyment of learning	Enjoyment of Learning (Pell and Jarvis's 2001)	–0.09 (–0.22, 0.03)
	Self-esteem	Global Self-Worth Scale of the Self-Perception Profile for Children (Harter, 1985)	–0.04 (–1.87, 0.10)
	Locus of control	Rotter's Locus of Control Scale	–0.05 (–0.31, 0.21)
Miller 2012	Enjoyment of reading	The Garfield Elementary Reading Attitudes Scale	0.03 (-0.11, 0.17)
	Reading confidence	The Reader Self-Perception Scale (Henk and Melnick, 1995)	0.03 (-0.13, 0.22)
	Aspirations for the future	Aspirations for the Future Scale (Loeber et al., 1991)	0.03 (-0.11, 0.17)

Table 5. Cochrane Collaboration Risk of Bias Tool application in the included studies

Study	Selection bias: sequence generation	Detection bias: blinding of outcome assessors	Attrition bias: incomplete outcome data
Allor 2004-1	?*	?	-*
Allor 2004-2	?	?	-
Baker 2000	?	+*	+
Cabezas 2011	?	?	+
Elliott 2000	?	?	+
Ham 1977	+	?	+
Jensen 1991	?	+	+
Lee 1980	?	?	-
Lee 2012	?	?	-
Loenen 1989	?	?	-
Miller 2009	+	+	-
Miller 2012	+	+	-
PSA 2007	?	?	-
Pullen 2004	?	+	-
Rimm-Kaufman 1998	?	+	?
Ritter 2000	+	?	-

**Note + low risk of bias - high risk of bias ? unclear risk of bias*

Figure 1. Flowchart of study selection adapted from PRISMA Statement (Moher et al. 2009)

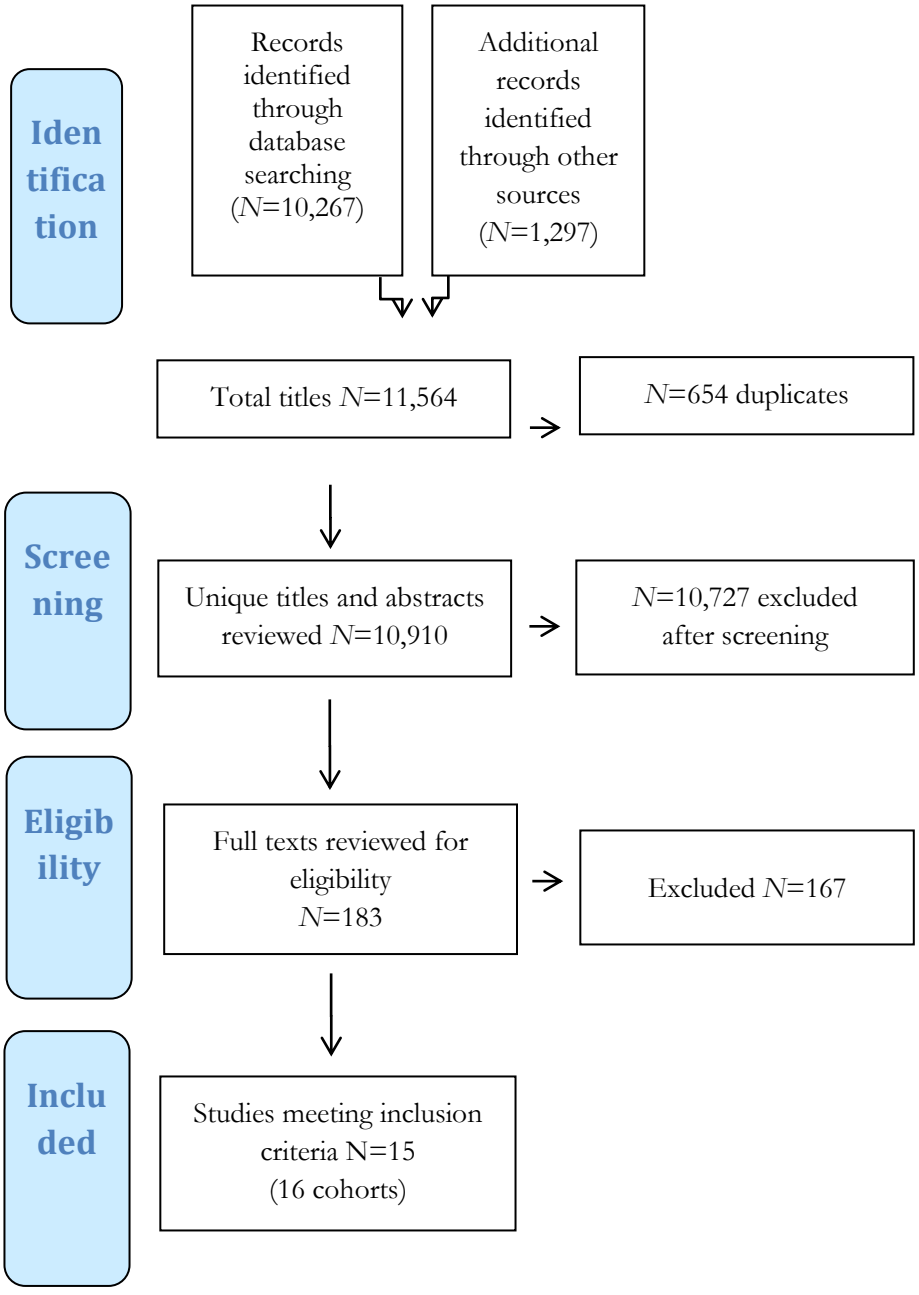


Figure 2. Forest plot of comparison between control group and tutoring on the composite measure of reading

Study name	Study size	Measure	Hedges's g	Lower limit	Upper limit	p-Value	Relative weight
Allor 2004-1	86	Combined	0.59	0.40	0.78	0.000	6.91
Allor 2004-2	157	Combined	0.46	0.34	0.58	0.000	8.34
Baker 2000	84	Combined	0.36	0.12	0.61	0.004	5.84
Cabezas 2011	4903	Combined	0.06	-0.01	0.13	0.102	9.13
Elliott 2000	30	Combined	-0.06	-0.55	0.43	0.806	2.69
Ham 1977	147	Combined	0.22	0.06	0.38	0.007	7.54
Jensen 1991	103	Combined	0.36	0.20	0.53	0.000	7.43
Lee 1980	40	SAT reading, gr. 3-6	0.06	-0.55	0.67	0.849	1.95
Lee 2012	883	Combined	0.09	0.01	0.16	0.028	9.07
Loenen 1989	81	Combined	-0.37	-0.68	-0.06	0.019	4.79
Miller 2009	734	Combined	0.09	-0.00	0.18	0.051	8.89
Miller 2012	483	Combined	0.10	0.03	0.17	0.007	9.12
PSA 2007	124	Terra nova reading	0.04	-0.31	0.39	0.833	4.16
Pullen 2004	47	Combined	0.61	0.20	1.02	0.004	3.47
Rimm-Kaufman 1998	42	Combined	0.18	-0.17	0.52	0.309	4.25
Ritter 2000	328	SAT-9 Reading	-0.11	-0.33	0.10	0.299	6.43
			0.18	0.08	0.27	0.000	

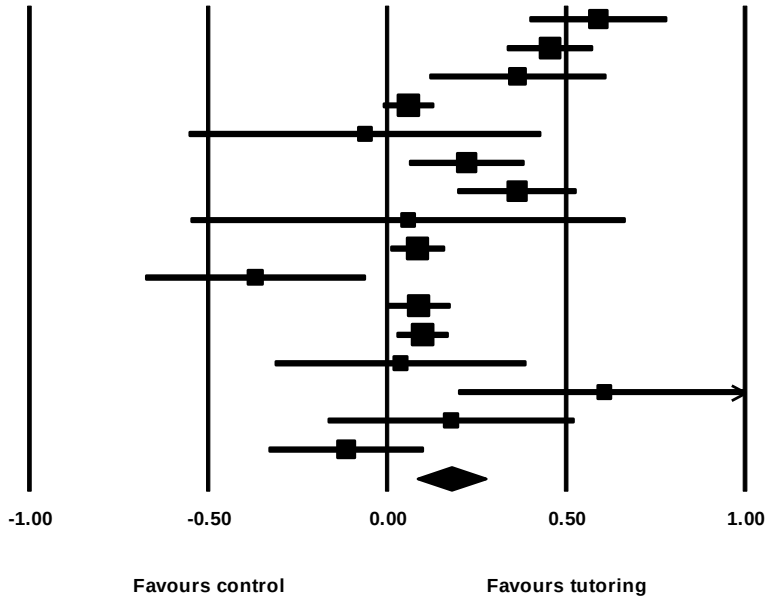


Figure 3 Funnel plot of standard errors by Hedges's g *for composite measure of reading, random-effects*